Personalisierte Audiodeskription mit KI-basierter Sprachsynthese

Bachelor-Thesis im Studiengang Audiovisuelle Medien an der Hochschule der Medien Stuttgart zur Erlangung des akademischen Grades „Bachelor of Engineering“.

Vorgelegt von:	Franziska Untraut
Matrikelnummer:	38487
Betreuer:	Prof. Dr. Gottfried Zimmermann (1. Prüfer)
	Sebastian Koch, MSc. (2.Prüfer)
Vorgelegt am:	27.03.2023

Ehrenwörtliche Erklärung

Hiermit versichere ich, Franziska Untraut, ehrenwörtlich, dass ich die vorliegende Bachelorarbeit mit dem Titel: „Personalisierte Audiodeskription mit KI-basierter Sprachsynthese“ selbstständig und ohne fremde Hilfe verfasst und keine anderen als die angegebenen Hilfsmittel benutzt habe. Die Stellen der Arbeit, die dem Wortlaut oder dem Sinn nach anderen Werken entnommen wurden, sind in jedem Fall unter Angabe der Quelle kenntlich gemacht. Die Arbeit ist noch nicht veröffentlicht oder in anderer Form als Prüfungsleistung vorgelegt worden.

Ich habe die Bedeutung der ehrenwörtlichen Versicherung und die prüfungsrechtlichen Folgen (§26 Abs. 2 Bachelor-SPO (6 Semester), § 24 Abs. 2 Bachelor-SPO (7 Semester), § 23 Abs. 2 Master-SPO (3 Semester) bzw. § 19 Abs. 2 Master-SPO (4 Semester und berufsbegleitend) der HdM) einer unrichtigen oder unvollständigen ehrenwörtlichen Versicherung zur Kenntnis genommen.

Stuttgart, den 27.03.23	F. Untraut
Ort, Datum	**Unterschrift**

Kurzfassung

Die Audiodeskription ermöglicht blinden und sehbeeinträchtigten Menschen das Verstehen und Erfahren von Videos und Filmen, indem visuelle Informationen in Dialogpausen beschrieben werden. Da die Anforderungen der Audiodeskription von Zielgruppe zu Zielgruppe stark variieren und auch der wirtschaftliche Faktor eine erhebliche Rolle spielt, existieren alternative Ansätze und Erweiterungen zur Erstellung von Audiodeskriptionen.

Neben den Grundlagen der Audiodeskription und der Sprachsynthese, beschäftigt sich die vorliegende Arbeit mit bereits vorhandenen Technologien zur Erweiterung und Vervielfältigung des Angebots von Audiodeskriptionen. Außerdem wird im praktischen Teil der Arbeit ein neuer Ansatz zur Personalisierung von Audiodeskriptionen vorgestellt. Dieser untersucht, ob ein Mehrwert geschaffen wird, indem die bisher mit menschlicher Stimme produzierten Audiodeskriptionen für Filme und Videos durch eine personalisierte künstliche Stimme ersetzt werden. Diese Personalisierung beinhaltet eine wählbare Sprechgeschwindigkeit, welche mit der Ausführlichkeit der Beschreibung zusammenhängt. Je höher die durch die Nutzenden bestimmte Geschwindigkeit, desto mehr Inhalt wird vermittelt. Es kann während des Abspielens zwischen drei verschiedenen Stufen gewechselt werden.

Der Ansatz wird in dieser Arbeit in einer Feldstudie anhand eines Prototyps getestet und bewertet.

Abstract

Audio description enables blind and visually impaired people to understand and experience videos and films by describing visual information in dialog pauses.

Since the requirements of audio description vary greatly from target group to target group and the economic factor also plays a significant role, there are alternative approaches and extensions for creating audio descriptions. Besides the basics of audio description and speech synthesis, this thesis deals with already existing technologies for the extension and increasing of the offer of audio descriptions. In addition, the practical part of the thesis presents a new approach to the personalization of audio descriptions. This investigates whether added value is created by replacing audio descriptions for movies and videos, which have been produced with human voices so far, with a personalized artificial voice. This personalization includes a selectable speaking rate, which is related to the verbosity of the description. The higher the speed determined by the user, the more content is conveyed. It is possible to switch between three different levels during playback.

This approach is tested and evaluated in a field study using a prototype.

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

AD ..... Audiodeskription

ADLAB ..... Audio Description: Lifelong Access for the blinds

BITV ..... Barrierefreie-Informationstechnik-Verordnung

CNNs ..... convolutional neural networks

CTS ..... context-to-speech

DAW ..... Digital Audio Workstation

IPA ..... International Phonetic Association

KI ..... künstliche Intelligenz

MeMAD ..... Methods for Managing Audiovisual Data

PSOLA ..... pitch synchronous overlap add

RNNs ..... recurrent neural networks

SSML ..... Speech Synthesis Markup Language

TTS ..... text-to-speech

WCAG ..... Web Content Accessibility Guidelines

1 Einleitung

Der Film- und Fernsehkonsum spielt im täglichen Leben von Menschen mit Sehbeeinträchtigung eine wichtige Rolle. Wenn dieser Teil fehlt, führt das nicht nur zu einer geringeren sozialen Interaktion, sondern kann sich auch negativ auf das Selbstbild und das Selbstvertrauen auswirken (Sackl et al., 2020). Daher ist ein barrierefreier oder mindestens ein barrierearmer Zugang zu Videoinhalten notwendig. Für blinde Menschen wird dieser durch das Beschreiben des visuellen Geschehens im Film, einer sogenannten Audiodeskription (AD), geschaffen. Weitere Zugänglichkeiten für Menschen mit Beeinträchtigungen werden durch die Verfügbarkeit von Untertiteln, eines Transkripts und eines Videos mit Gebärdensprache erreicht.

Orero (2022, S.407f.) behauptet, dass sich der Auftrag der Barrierefreiheit für Menschen mit Behinderung vom ausschließlichen Ansatz einer Zugänglichkeit entfernt und vielmehr den Fokus auf den Bereich des Individuums und seine Fähigkeiten verlagert. Sie betont: „Jeder einzelne Nutzer hat sein eigenes Profil an Bedürfnissen, Eigenschaften, Fähigkeiten und Vorlieben. Diese Tatsache muss bei der Entwicklung von Mainstream-Produkten und Dienstleistungen berücksichtigt werden“ (Orero, S. 407f.).

Eine Zunahme der Personalisierung von digitalen Medienerlebnissen wird von Evans et al. (2022) bestätigt. Sie ergänzen, dass nicht nur Webseiten und Werbungen individualisiert werden, sondern auch die Erfahrungen des Video-Streamings inzwischen stark unterschiedlich sind (Evans et al., 2022).

Früher wurden die „linearen“ Medien als einzige Spur erkannt und nur die Lautstärke, die Helligkeit und der Kontrast konnten eingestellt werden. Seit Medien als unabhängige Einheiten mit Bild, Ton und Text produziert und aufgezeichnet werden, ist die Personalisierung von Videos möglich. Eine individuelle Kombination aus Zugänglichkeitsdiensten durch die Wahl der Nutzenden wurde dadurch ermöglicht (Orero, 2022, 407ff.).

Untertitel können bereits häufig der Textgröße, der Schriftart, der Hintergrundfarbe und der Positionierung im Bild angepasst werden. Die Möglichkeiten zur Individualisierung einer Audiodeskription beziehen sich auf das Einstellen der Sprechgeschwindigkeit, der Sprecherstimme und der Tonmischung, sowie das Entscheiden über den Inhalt der Beschreibung (Orero, 2022, S.410). Diese Optionen werden jedoch von den derzeit verbreiteten Playern noch nicht bereitgestellt.

Die vorliegende Arbeit beschäftigt sich mit diesem Thema und schlägt mithilfe neuester Technologie ein neues Modell der personalisierten Audiodeskription vor.

2 Methodik

In diesem Kapitel wird das methodische Vorgehen beschrieben, welches zur Datenerhebung und Datenanalyse eingesetzt wurde, um Aussagen über die Forschungsfrage treffen zu können.

2.1 Ziel der Arbeit

Das zu erreichende Ziel der vorliegenden Arbeit ist es, anhand vorheriger Recherche zum Thema *Audiodeskription* und *Sprachsynthese* ein Modell der personalisierten Audiodeskription zu erstellen, bei dem die Sprechgeschwindigkeit und dementsprechend auch die Ausführlichkeit des Inhalts variierbar ist. Dieses Modell wird mithilfe von Sprachsynthese realisiert und anhand eines Usability-Tests, an dem sechs blinde Menschen teilnehmen, erprobt und bewertet.

Es gilt, diesen Ansatz ebenfalls in Bezug auf verschiedene Filmgenres zu untersuchen, um herauszufinden, ob sich das System für die verschiedenen Genres besser oder schlechter eignet. Außerdem wird ein Vergleich zwischen dem Modell mit Sprachsynthese und einem Standardsystem mit menschlicher Stimme gezogen.

Die Hauptfrage besteht darin, herauszufinden, ob die Eigenschaft der Emotionsübertragung einer menschlichen Stimme oder die Flexibilität einer künstlichen Stimme von den Nutzenden bevorzugt wird. Mit den Ergebnissen der Studie soll abschließend herausgefunden werden, in welchen Bereichen es einer Verbesserung bedarf, damit diese Art von personalisierter Audiodeskription als Bereicherung in den Alltag von sehbeeinträchtigten Menschen eingebunden werden kann.

2.2 Vorgehen

Um das obengenannte Ziel zu erreichen, wird in Kapitel 3 der Arbeit ein grundlegendes Verständnis für die Erstellung einer Audiodeskription (AD) geschaffen. Das Verständnis wird durch das Erläutern von Methoden und Techniken der Filmbeschreibung, das Aufzeigen ihrer Einsatzgebiete und durch die Beachtung verschiedener Genres vermittelt. Außerdem wird die Rolle der Audiodeskription in der Emotionsvermittlung analysiert und ihr Stellenwert diesbezüglich abgeleitet.

In Kapitel 4 wird die Funktionsweise der Synthese von Text in Sprache erklärt und die künstliche Stimme mit der menschlichen Stimme verglichen. Dabei werden Vor- und Nachteile des

Einsatzes von synthetisierten Stimmen in der Audiodeskription gegenübergestellt und die Schwierigkeiten der Synthese thematisiert.

Kapitel 5 widmet sich dem aktuellen technischen Stand im Bereich der Audiodeskription und der Sprachsynthese. Es werden neuartige Technologien zur automatischen Generierung eines AD-Skripts erläutert und verschiedene Erweiterungen der klassischen Audiodeskription vorgestellt. Um Erweiterungen jeglicher Art einbinden zu können, ist es erforderlich, dass die entsprechenden Videoplayer für eine breite Masse zugänglich sind. Daher werden die aktuellen Möglichkeiten bezüglich der Zugänglichkeit und Personalisierung von Videoplayern in der vorliegenden Arbeit anhand von drei Playern ermittelt und verglichen.

Der daran anschließende praktische Teil befasst sich mit der Feldstudie zum oben beschriebenen Ansatz der personalisierten Audiodeskription mit künstlicher Stimme und variierbarer Sprechgeschwindigkeit. Die dafür notwendige Konzipierung und Umsetzung eines Prototyps wird erklärt, sowie die Videowahl hinsichtlich der Filmgenres begründet. Außerdem werden die verwendeten Methoden des Usability-Tests aufgezeigt und die Studienpopulation beschrieben. Anschließend erfolgt eine Schilderung der Durchführung des Tests und die dazugehörige Auswertung der Ergebnisse.

2.3 Arbeitsweise

Die Idee des neuen Ansatzes basiert auf der 2021 verfassten Master-Arbeit von Remo Schneider. Im Rahmen dieser Thesis hat er den barrierefreien Able-Player entwickelt, der unter anderem die Möglichkeit einer erweiterten Audiodeskription besitzt. Diese beinhaltet die Wahl zwischen einer kurzen, mittellangen und ausführlichen Beschreibung durch den User. Da der Player mittels Text-To-Speech Technologie umgesetzt wurde, ist eine separate Einstellung der Sprechgeschwindigkeit der künstlichen Stimme möglich (Schneider, 2021).

Wir haben diesen Ansatz weiterentwickelt, indem wir die Sprechgeschwindigkeit an die Inhaltsmenge gekoppelt haben. Daraus resultiert eine Abhängigkeit der beiden Bestandteile, die unserer Meinung nach Sinn ergibt und in der Studie ebenfalls untersucht wird.

Um fundierte und präzise Aussagen über die Forschungsfragen treffen zu können, wurde zu Beginn Material aus bestehender Literatur zusammengetragen und analysiert. Außerdem ist durch die Rücksprache mit Prof. Dr. Zimmermann und Sebastian Koch, sowie durch die Befragung eines Autors von AD-Skripten, umfangreiches Expertenwissen in die Arbeit eingeflossen.

Der nächste Schritt war die Konzipierung des Prototyps, in welcher wir die Geschwindigkeitswahl auf drei Stufen begrenzt und eine Auswahl von drei verschiedenen Genres getroffen haben. Somit wurden Überschaubarkeit und Einschränkung des Umfangs geschaffen.

Um die Videos den Testpersonen zur Verfügung zu stellen, haben wir eine Einbindung des Players in eine Webseite für sinnvoll erachtet. Der Zugriff konnte dadurch ortsunabhängig erfolgen und eine Online-Teilnahme an dem Test wurde ermöglicht. Dementsprechend konnten wir die Suche nach Testpersonen auf ganz Deutschland ausweiten. Außerdem bietet eine Webseite die Option einer eigenständigen Bedienung des Players mittels Screenreader.

Für die Programmierung des Prototyps haben wir zunächst die standardmäßige HTML5-Funktion „AudioTracks“, die von video.js unterstützt wird, in den Code eingebunden. Damit wird das Einfügen und Wechseln mehrerer Audiospuren zu einem Video möglich gemacht. Anwendung findet diese Funktion bei Videos, die in verschiedenen Sprachen verfügbar sind. Für das System der erweiterten Audiodeskription wäre sie ebenfalls hilfreich gewesen, jedoch werden „AudioTracks“ nur von wenigen Browsern unterstützt. Aufgrund dessen haben wir uns gegen diesen Ansatz entschieden. Die weitere Überlegung, den Able-Player als Ausgangsplayer zu verwenden und diesen zu erweitern, haben wir aufgrund der Komplexität und anderweitiger Probleme verworfen. Die dritte Methode hat schließlich unseren Vorstellungen entsprochen. Sie besteht aus einem mittels JavaScript erweiterten HTML5-Videoplayer, welcher die Möglichkeit eines Videowechsels bietet. Vorgestellt wird die Programmierung in Kapitel 6.2.1.

Die Wahl der Filmausschnitte begründet sich in dem Versuch, möglichst gegenteilige Genres in die Forschung einzubeziehen, die die typischen Merkmale der entsprechenden Filmsorte repräsentieren. Somit ist die Abdeckung eines breiten Spektrums von Genres erreicht und die Chance auf klare Ergebnisse durch das Herausstechen von spezifischen Merkmalen erhöht.

In die Testung des neuen Systems der Audiodeskription wurden ausschließlich blinde Personen einbezogen, da diese die potenziellen Nutzerinnen und Nutzer sind und am besten einschätzen können, ob das dargelegte System ihren Alltag bereichern würde. Als Studienpopulation haben wir sechs Personen ausgewählt, weil diese Anzahl an Personen ein aussagekräftiges Ergebnis ermöglicht und der Test so im vorgegeben Zeitrahmen umsetzbar ist.

Um eine Zuverlässigkeit (reliability) des Tests und eine Repräsentation der Gesamtpopulation zu gewährleisten, wurden die Testpersonen hinsichtlich des Alters, des Geschlechts und des Zeitpunkts der Erblindung möglichst divers ausgewählt.

Da es sich bei der Forschung um eine qualitative Studie handelt und die Bewertung der Testpersonen subjektiv erfolgt, besteht ebenfalls eine Gültigkeit (validity) des Tests. Durch das Einsetzen von Counterbalancing wurde außerdem eine mögliche Auswirkung der Reihenfolge auf die Ergebnisse ausgeschlossen.

Die Objektivität (objectivity) wird aufgrund der Betreuung und Beobachtung der Testpersonen garantiert. Außerdem haben die Teilnehmenden den Test unabhängig voneinander durchgeführt.

In der Studie wurde das Design des „within-subjects“ angewendet. Durch diese Methode kann der Einfluss von Variablen, wie beispielsweise des Alters und des Geschlechts, reduziert werden, da eine Versuchsperson beide Systeme testet (Kuniavsky, 2003).

Als Methode zur Datenerhebung haben wir mit den einzelnen Testpersonen Interviews geführt, weil sich eine mündliche Befragung mit blinden Menschen einfacher gestaltet als das Ausfüllen eines schriftlichen Fragebogens. Außerdem steht mehr Platz für Begründungen und Erklärungen zur Verfügung, die bei dem eingesetzten Prinzip der „self-reported-metric“ von großem Nutzen sind, um die subjektiven Einschätzungen erfassen zu können. Die Fragen bezogen sich vor allem auf die Bewertung der Geschwindigkeitsstufen, auf die Eignung der verschiedenen Genres und auf die Beurteilung der künstlichen Stimme, wobei zur Beantwortung der Fragen häufig eine Likert-Skala hinzugezogen wurde. Aus den Ergebnissen dieser Skalen wurden statistische Analysen durchgeführt, um einen möglichen Mehrwert der personalisierten Audiodeskription zu untersuchen.

3 Audiodeskription in Filmen und Videos

Dieses Kapitel vermittelt Daten und Fakten zum Thema der Audiodeskription und beschreibt grundlegende Techniken zum Erstellen eines AD-Skripts. Außerdem wird die Funktion der Audiodeskription im Hinblick auf spezifische Filmgenres und bezüglich des Emotionstransfers untersucht. Die verschiedenen Nutzergruppen von Audiobeschreibungen werden ebenfalls in diesem Kapitel betrachtet.

3.1 Einsatzgebiete der Audiodeskription

Audiodeskriptionen werden hauptsächlich bei Filmen und Serien im öffentlichen Fernsehen zur Verfügung gestellt, da die Verordnung zur Schaffung barrierefreier Informationstechnik (BITV) nach dem Behindertengleichstellungsgesetz (BGG) eine Zugänglichkeit für Menschen mit Beeinträchtigung zu audiovisuellen Inhalten von öffentlich-rechtlichen Anbietern vorschreibt. Eine Mindestanzahl von Sendungen und Filmen ist jedoch nicht vorgegeben (*BITV 2.0*, o. J.).

Seit 2013 werden im Hauptprogramm des Ersten alle fiktionalen Formate, d.h. Krimis, Spielfilme und Serien, sowie Tier- und Naturfilme, in einer Hörfilmfassung ausgestrahlt (Heerdegen-Wessel, 2019, S.731).

In folgender Abbildung 1 kann die stetig ansteigende Zahl von verfügbaren Audiodeskriptionen im NDR-Fernsehen feststellt werden.

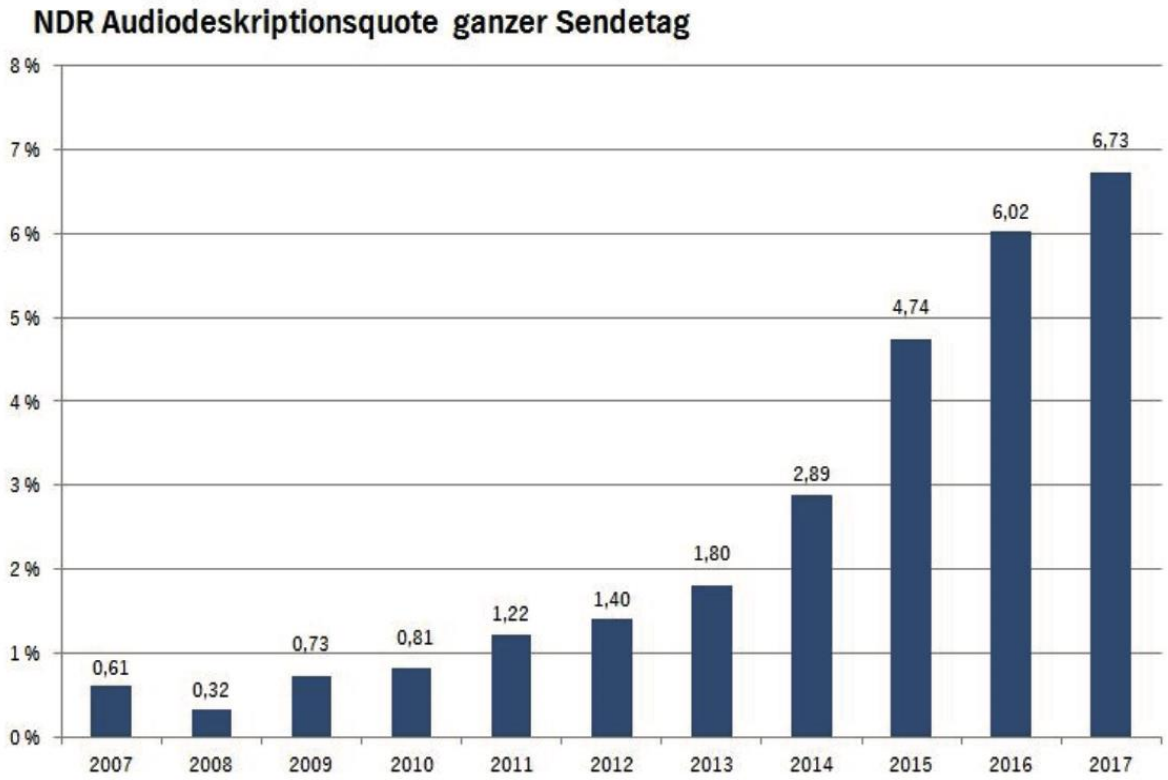

Abbildung 1: NDR Audiodeskriptionsquote ganzer Sendetag (Heerdegen-Wessel, 2019, S.731)

Das private Fernsehen hingegen stellt nur selten Hörfilmfassungen für Sendungen und Filme bereit. Auf DVDs, sowie bei Filmen auf den immer wichtiger werdenden Streaming-Plattformen sind häufig Hörfilmfassungen vorhanden, die laut Schruhl (2019, S.776ff.) jedoch unterschiedlich gut bedienbar sind und teilweise zusätzlich kosten.

Damit sehbeeinträchtigte Menschen bei Kinofilmen ebenfalls eine Audiodeskription erhalten können, bieten manche Kinos die Möglichkeit, Empfangsgeräte auszuleihen, wodurch Filmbeschreibungen über bestimmte Sitzplätze abgerufen werden können. Voraussetzung hierfür ist eine vorhandene Hörfilmfassung (Schruhl, 2019, S.776ff). Inzwischen gibt es außerdem die Möglichkeit eine App für Audiodeskriptionen zu verwenden, welche in Kapitel 5.3 näher erklärt wird.

Die Anwendungsgebiete von Audiodeskriptionen sind jedoch viel weitreichender. Sie reichen von Museen und Galerien über Bildungseinrichtungen und architektonische Führungen, bis hin zu Beschreibungen von Sportevents und Live-Veranstaltungen (Fryer, 2016, S.1; Snyder, 2005, S.1ff).

In Museen sorgt die Audiodeskription für eine lebendige und fantasievolle Führung, wodurch nicht nur für sehbeeinträchtigte Menschen, sondern für alle Besucher eine bereichernde Wirkung erreicht werden kann. Manchmal wird eine zusätzliche Version zu einer bestehenden Audiotour hinzugefügt, die Richtungsinformationen oder Filmbeschreibungen von Filmen

innerhalb der Ausstellung enthält, damit Blinde einen Teil des Museums selbstständig besichtigen können.

Des Weiteren werden Audiodeskriptionen in Bildungseinrichtungen beim Lesen von Kinderbüchern eingesetzt, um ebenfalls die Sprachkenntnisse von sehbeeinträchtigten Kindern zu fördern (Snyder, 2005, S.938).

Snyder (2005, S.937) und Walczak (2018, S.833) sind sich einig, dass Audiodeskriptionen nicht nur bei blinden und sehbeeinträchtigten Menschen Anwendung finden, sondern auch für sehende Menschen aus praktikablen Gründen einen Mehrwert erzeugen können. Damit ist die Verwendung von Audiodeskriptionen bei Aktivitäten wie beispielsweise Autofahren oder Abwaschen gemeint, bei denen es nicht möglich ist, auf das entsprechende Endgerät zu schauen.

3.2 Methoden und Techniken der Filmbeschreibung

Eine Audiodeskription erfordert eine geeignete Verwendung von Sprachmitteln, sowie eine sinnvolle Wahl der zu beschreibenden Elemente im Film. Hierfür gibt es mehrere Methoden und Techniken, die in diesem Kapitel vorgestellt werden.

Die Ausarbeitung wird von Sehenden und Blinden gemeinsam bewältigt, wobei sich das Team meist aus ein oder zwei sehenden Beschreibenden und einem nicht sehenden Teammitglied zusammensetzt. Die Positionen verteilen sich auf eine navigierende Person, die den Film steuert, einer Texterin oder einem Texter, der das Skript dokumentiert und einer sehbeeinträchtigen Person. Zu den Hauptaufgaben dieser sehbeeinträchtigten Person gehören die Sequenzierung und die Selektion von zu beschreibenden Informationen, sowie das Erkennen und Lösen von Übersetzungsproblemen. Unumgänglich ist vor allem das Einschätzen der Adäquatheit des Deskriptionstextes durch die nichtsehende Person (Hirvonen & Schmitt, 2018, S.9). Der Hörfilm e.V. kritisiert, dass aufgrund von Kosteneinsparungen immer häufiger Einzelpersonen beauftragt werden *(Hoerfilm e.V. | Audiodeskription,* o. J.).

Die Beschreibung des Filmes geschieht normalerweise in den Dialogpausen, wobei die Beachtung der Soundeffekte und der Musik nicht vernachlässigt werden sollten (Benecke, 2019, S.455). Fachleute sind sich dem entgegen nicht einig, wenn es sich um den Inhalt und die Art und Weise des Beschreibens handelt. Es gibt zwar Regelwerke, die als Leitfaden dienen und in Europa weitgehendst gleich sind, jedoch ist Benecke (2019, S.458) der Meinung, dass diese Regelwerke zu große Spielräume aufweisen und sich daher die Stile der Audiodeskriptionen stark unterscheiden. Beispiele für Regelwerke sind die *ITC Guidance on Standards for Audio Decription* in Großbritannien, die *Requisitos para la audiodescripción* in Spanien und die *audio description guidelines* in Deutschland. Letztere wurden 2015 aus einer Initiative von deutschen Rundfunkanbietern in Kooperation mit deutschen Hörfilm-Vereinen erstellt (*NDR*,

o. J.). Länderübergreifend gibt es in Europa seit 2014 die ADLAB-Richtlinien, die in Kooperation vieler europäischer Länder erschaffen wurden.

Um die wichtigsten Informationen in die Audiodeskription einzuschließen, wird dort empfohlen, den Ort, die Zeit, die beteiligten Personen, die wichtigsten Gegenstände bzw. Gebäude und die für das Verständnis notwendige Handlung in die Beschreibung einzubeziehen.

Bei den Personenbeschreibungen ist es wichtig, dass sie im Film so früh wie möglich einen Platz finden. Dem pflichtet auch Benecke (2019, S.460) bei und ergänzt, dass der Zeitpunkt der Namensgebung der Figur jedoch mit Bedacht gewählt werden sollte. Denn der Narration darf durch die Audiodeskription nichts vorweggenommen werden. Diese Gefahr besteht ebenfalls in der Detailliertheit der Beschreibung einer Person. Hauptpersonen und wichtige Personen für die Handlung werden detaillierter beschrieben als andere. Diese Differenziertheit lässt die Sehbeeinträchtigten auf die Bedeutsamkeit der Person rückschließen. Die Art der Benennung einer Person ist ebenfalls maßgeblich. Wenn die Person mit dem Vornamen bezeichnet wird, wird dem Zuhörer eine Nähe vermittelt und eine jüngere Person wird erwartet. Wohingegen die Bezeichnung mit dem Nachnamen und möglicherweise sogar mit einer Anrede, eine Distanz erzeugt und eine ältere Person erwarten lässt. Eine noch größere Distanz und somit eine geringe Identifikation wird durch allgemeine Bezeichnungen, wie „ein Mensch“ oder „ein Junge“ erzeugt (Hämmer, 2005, S.87f und S.97).

Bei der Handlungsbeschreibung einer Person sollte vor allem Wert auf Gestik, Mimik und Blickrichtung der betreffenden Person gelegt werden (Jüngst, 2020, S.115).

Da die Dialogpausen häufig kurz sind, ist oft wenig Platz für die Audiobeschreibungen. Aus diesem Grund müssen sie möglichst informativ, knapp und ausdrucksstark formuliert sein. Es gibt einige sprachliche Mittel, womit das erreicht werden kann.

Hierzu zählt eine einfache Syntax mit zum Teil unvollständigen Konstruktionen. Ein Beispiel für solche Ellipsen ist: „Ein Parkplatz.“ anstatt „Die nächste Szene spielt auf einem Parkplatz.“ Zeit- und Ortsangaben erfolgen grundsätzlich am Beginn eines Szenenwechsels (Fix, 2005, S.144). Außerdem werden Partizipien mit adverbialer Funktion und starke Attribuierungen verwendet. Hierzu zählen zum Beispiel die Partizipien „überquellend“, „verschwommen“, „kopfschüttelnd“. Die Umstände werden hiermit genau beschrieben und erzeugen ein präzises Bild im Kopf des Zuhörers. Um eine sprachliche Kompression zu erzeugen, können Partizipien auch hintereinander gesetzt werden: „Dreißigjährig, mittelblond, blauäugig und blutgetränkt“(Poethe, 2005, S.44). Die Verwendung von Verben, die sich auf bestimmte Arten, etwas zu tun, beziehen, sind ebenfalls förderlich. Der Ausdruck „sie tänzelt/marschiert/…“ anstatt „sie geht“ verbessert die Aussage deutlich (Salway, 2007, S.6).

Um weitere Verknappungen zu erzielen ist die Substantivierung von Verben (bspw. „beim Hinausgehen“) und der Einsatz substantivischer Komposita (bspw. „Treppenhausausgang“) ein beliebtes Mittel. Selbst substantivische Konversionen wie „die Tote“, „der Schnauzbärtige“ und „der Dicke“ sind ein effektiver Lösungsweg zur Identifizierung und zur Veranschaulichung einer Person. Zu den sprachlichen Eigenschaften eines Audiodeskriptionstextes ist hinzuzufügen, dass im Wortschatz keine umgangssprachlichen Wörter wie „ne“, „also“ oder „naja“ vorhanden sind und dass das verwendete Tempus das Präsens ist.

Daraus lässt sich schließen, dass Audiodeskriptionen Texte mit eigenen Gesetzen sind (Morgner & Steffen Pappert, 2005; Poethe, 2005). Die Studie von Salway (2007) bestätigt diese Aussage. Seiner Ansicht nach ist die Audiodeskription sogar eine „spezielle Sprache“ (Salway, 2007, S.31).

Um dieser eigenen Sprache noch tiefer auf den Grund zu gehen, bedarf es einer spezifischen Filmanalyse. Es gibt verschiedene Modelle, von denen zwei im Folgenden genauer vorgestellt werden.

Mazur (2020) schlägt für die erfolgreiche Erstellung einer Audiodeskription eine Analyse des Ausgangsprodukts, also des Films ohne Audiodeskription, auf drei Ebenen vor: *die kontextuelle Ebene, die makrotextuelle Ebene* und *die mikrotextuelle Ebene*.

Auf der *kontextuellen Ebene* werden grundlegende Faktoren analysiert. Es wird untersucht an wen sich der Film richtet, zu welcher Zeit und an welchem Ort er konsumiert wird, auf welchem Medium er gezeigt wird (on-demand, TV, Festival,…), zu welchem Genre der Film gehört und welche Absicht die Audiodeskription verfolgt. Mazur merkt an, dass das Ausgangsprodukt auf Multimodalität basiert, also unterschiedliche semiotische Zeichensysteme wie Sprache, Bilder und Geräusche, besitzt. Daher sei also eine multimodale Analyse erforderlich.

Aus diesem Grund wird anschließend in der *makrotextuellen Analyse* ermittelt, welche Information sowohl in Dialogen, in der Musik als auch im Sound bereits übermittelt werden. Dem pflichtet auch Fryer bei. Sie legt einen Fall dar, in dem die Übermittlung der Soundeffekte nicht berücksichtigt wurde. Es wird dort beschrieben: „There is a knock at the door, and you hear ‘bang bang bang’“ (Fryer, 2016, S.29). Sie vertritt die Meinung, es sollte lieber beschrieben werden, wer an der Tür ist, und nicht das Offensichtliche. Darüber hinaus werden in der Makroanalyse die rahmengebenden Faktoren des Inhalts und der Struktur analysiert. Hierzu zählt die Analyse des kulturellen Hintergrunds, das Untersuchen der Charaktere und deren Beziehungen, eine räumliche und zeitliche Betrachtung, sowie die Interpretation des Titels. Ebenfalls wird der Stil der Sprache in Bezug auf das Genre untersucht. Dies wird in Kapitel 3.4 näher erläutert.

Wie der Name der *mikrotextuellen Ebene* schon erahnen lässt, werden hier spezifische Szenen oder sogar einzelne Einstellungen analysiert. Komplexe visuelle Szenen werden auf einige der gleichen Parameter, wie in der Makroanalyse (Charaktere, Sprache, Geräusche, Zeit, Raum) untersucht und es wird entschieden, welche Elemente auf niedriger Ebene für die Audiodeskription wichtig sind (Mazur, 2020).

Eine andere Methode um das komplexe Konstrukt eines Film herunterzubrechen, ist die wissenschaftliche Filmanalyse mit den Prinzipien von Henrike Morgner und Steffen Pappert (Morgner & Steffen Pappert, 2005). Hierbei werden Sequenzprotokolle und Einstellungsanalysen als unterstützende Arbeitsmittel eingesetzt.

Im ersten Schritt wird der Film in Sequenzen, sogenannte Handlungseinheiten, die beispielsweise durch einen Ortswechsel markiert werden, eingeteilt. Danach findet die verfeinernde Einteilung in Subsequenzen und Einstellungen zur Erfassung der filmischen Struktur statt. Der gesamte Aufbau wird im Sequenzprotokoll festgehalten.

Im zweiten Schritt wird basierend auf dem Sequenzprotokoll ein Einstellungsprotokoll angefertigt, welches die filmästhetischen Merkmale erfasst. Es wird in der untenstehenden Abbildung 2 in einer Tabelle dargestellt und beinhaltet die Bildnummer, die Dauer des Bildes, eine visuelle Beschreibung, die Einstellungsgröße, die Kamerabewegung, den Kamerastandort und die Kameraperspektive.

„a)“ bezeichnet hier eine Subsequenz, die aus drei Einstellungen bzw. Bildern besteht.

Sequenz 1 (Kommissariat Dresden)

Bild	Dauer	Visuelle Beschreibung	Einstellungsgröße	Kamerabewegung	Kamerastandort	Kameraperspektive
1	8 Sek.	**a)** Außen, eine Kindergruppe (Schulklasse) geht in Zweierreihe über eine Straße, im Hintergrund sieht man die Ruine der Frauenkirche und zwei Kräne; die Kindergruppe bleibt an einem roten Motorrad stehen, das am Straßenrand geparkt ist	Halbtotale (HT)	Still, dann Schwenk nach unten	Auf dem Bürgersteig vor dem Motorrad	Brusthöhe
2	4 Sek.	Einige Kinder aus der Gruppe gehen um das rote Motorrad herum, die Erzieherin in einem rot karierten Hemd mahnt die Kinder, nicht stehen zu bleiben	Nah	-	Auf der anderen Seite des Motorrads	Von unten
3	7 Sek.	Man sieht die Kinder, die das Motorrad noch näher betrachten, von oben, einige fassen es an	Totale (T)	-	Am Fenster im Gebäude	Obersicht

Abbildung 2: Auszug aus der Einstellungsanalyse "Laura, mein Engel", Teil der Sequenz 1 (Morgner & Steffen Pappert, 2005)

Optional kann eine Spalte für die Dokumentierung der Dialoge hinzugefügt werden.

Neben den Informationen, die bereits durch die Soundeffekte und die Musik übermittelt werden, muss das sogenannte Schemawissen berücksichtigt werden. Hierbei handelt es sich um bereits vorhandenes Wissen der Nutzer zu charakteristischen Handlungsabläufen. Alltägliche Situationen und Umgebungen, wie das Einkaufen im Supermarkt, müssen nicht so genau beschrieben werden, wie unübliche Situationen, da die Handlungsabläufe bereits bekannt sind. Jedoch muss auch berücksichtigt werden, dass Rezipienten unterschiedliche Schemata haben können. Ein Beispiel hierfür ist, dass in Südwest-England Scheunen aus Stein sind, während sie in Ost-England aus Holz sind (Fryer, 2016; Jüngst, 2020, S.115; Yos, 2005, S.102).

Abschließend lässt sich sagen, dass in einer Szene viele Elemente um die Aufmerksamkeit des Zuschauers konkurrieren. Sehende Personen können die Szene ganzheitlich aufnehmen, wohingegen eine Audiodeskription für blinde Menschen linear erfolgt.

Daher müssen die wichtigsten Elemente ausgewählt werden. Hierfür existieren die genannten Analysen und Methoden zur Hilfestellung, jedoch ist ein gewisser Teil immer subjektiv.

Vor allem die Menge der Informationen ist nicht klar definiert. Während britische Verfassende jede Lücke nutzen und sogar gelegentlich während des Dialogs beschreiben, wird in Amerika auf weniger Text und mehr Platz für Musik und Soundeffekte geachtet (Poethe, 2005, S.35).

Die gewünschte Menge von Informationen unterscheidet sich von Zielgruppe zu Zielgruppe, worauf im nächsten Kapitel eingegangen wird. Auch von Person zu Person bestehen individuelle Wünsche, wodurch weiterführende Möglichkeiten zur Personalisierung gefragt sind.

3.3 Zielgruppe von Audiodeskriptionen

Bei Weitem nicht alle AD-Nutzenden sind komplett blind. In Deutschland leben 155.000 blinde und 500.000 sehbeeinträchtigte Menschen (Anishchenko, 2020, S.23). Da jedoch keine Meldepflicht besteht, wird die reale Zahl auf etwa 1,2 Millionen Menschen geschätzt, und die Prognosen vermuten einen weiteren Anstieg (Anishchenko, 2020, S.23). Eine Vielzahl davon sind ältere Menschen, die Hör- und Sehverluste aufgrund des Alters erfahren. In England sind von 143.385 blind registrierten Menschen 61 Prozent über 75 Jahre alt (Fryer, 2016, S.42f). In Bayern sind Stand Oktober 2020 40,3 Prozent der Menschen, die Blindengeld erhalten, über 80 Jahre alt (*Zahlen & Fakten zu Blindheit und Sehbehinderung, o. J.).*

Während anzunehmen ist, dass Menschen ab einem gewissen Alter eine langsamere Beschreibung mit weniger Inhalt benötigen, können Blinde hingegen Aufgaben, bei denen eine auditive Aufmerksamkeit gefordert ist, sehr gut bewältigen. Das oft besser als sehende Menschen (Matamala & Orero, 2016, S.145). Die Zielgruppen unterscheiden sich jedoch nicht nur im Alter. Auch innerhalb von Gruppen sehbeeinträchtigter Menschen entstehen aufgrund abweichender Erfahrungen und unterschiedlichen Seherkrankungen differenzierte Anforderungen an die Audiodeskription. Kuppusamy und Pantula (2019, S.4f) berichten von zwei Nutzern mit gegenteiligen Bedürfnissen. Der eine Nutzer hat erst kürzlich sein Augenlicht verloren, weshalb ihn vor allem die Emotionen, der Tonfall und die Darstellung der Figur interessieren. Wohingegen der andere schon sein Leben lang blind ist und anstatt Beschreibungen von Farben, überwiegend die Schlüsselaspekte erfahren möchte.

Matamala und Orero (2016, S.59) erklären, dass das „geistige Bild“ von Menschen, die von Geburt an blind sind, ein anderes ist, als das von sehenden Menschen. Bei Blinden ist die mentale Vorstellung stark räumlich geprägt und von Beweglichkeit und Taktilität abhängig.

Außerdem ist das Schemawissen, also das Vorwissen aus Erfahrungen ein anderes. Manchen geburtsblinden Menschen wurde beispielsweise nie erklärt, wie es aussieht, wenn Menschen nicken oder Anführungszeichen in die Luft machen (Matamala & Orero, 2016, S.62).

Die Anforderungen an die Audiodeskriptionen variieren also stark von Zielgruppe zu Zielgruppe, sodass Jüngst (2020, S.106) zurecht feststellt, dass zwangsweise Kompromisse eingegangen werden müssen, um eine zufriedenstellende Audiodeskription für eine möglichst große Schnittmenge zu produzieren. Jedoch hat Salway (2007, S.27) erkannt, dass die Kompromisse nicht allzu groß sein müssten, wenn man die Möglichkeit zu einer Anpassung der Audiodeskription hätte. Sein erster Vorschlag ist das Bereitstellen verschiedener Beschreibungsversionen, die sich mit unterschiedlichem Vokabular an das Alter anpassen, sowie die Präferenz für beschreibende bzw. interpretierende Informationen berücksichtigen.

Seine zweite Idee ist das optionale Pausieren des Films, um detailliertere Informationen zu erhalten. Dieser und weitere Ansätze zur Personalisierung von Audiodeskriptionen werden in Kapitel 5.3 ausgeführt.

3.4 Audiodeskription in Bezug auf Filmgenres

Das Team, welches 2011-2014 die ADLAB-Leitlinien konzipierte, beschrieb wie schwierig es ist, starre Regeln für die Erstellung von Audiodeskriptionstexten zu entwerfen. Der Hauptgrund für die Schwierigkeiten war das breite Spektrum von Genres. Für die Erstellung einer Audiodeskription ist es sinnvoll, genrespezifische Merkmale zu beachten, da sie für die Prioritätensetzung der wichtigen Elemente, die in die Beschreibung eingebunden werden, ausschlaggebend sind (*ADLAB Audio Description guideline,* o. J.).

Im Folgenden werden einige Genres mit ihren spezifischen Merkmalen bezüglich Formalität, Ästhetik und Narration aufgezeigt. Zudem wird die dazugehörige Bedeutung für die Audiodeskription erklärt (*ADLAB Audio Description guideline,* o. J.; Fryer, 2016).

<u>Komödie:</u>

Eine Komödie ist häufig hell beleuchtet, um die heitere Atmosphäre des Films zu unterstützen. Die Lichtstimmung sollte also in der Beschreibung erwähnt werden.

Oft wird der Humor einer Komödie in der Audiodeskription nicht ausschließlich in den Vordergrund gestellt. Tyfour (2021, S.33) erklärt in seinem Buch, dass Martìnet-Sierra den Film „I want Candy“ auf die Beschreibung des Humors in Komödien untersuchte und auf das Ergebnis kam, dass dort 40% des visuellen Humors nicht beschrieben wurden.

Actionfilm:

In einem Actionfilm ist der Gebrauch von schnellen Bildwechseln üblich, um das Tempo der Ereignisse zu erhöhen.

Das bedeutet, dass wenig Zeit für die Audiodeskription zur Verfügung steht und diese umso kürzer und ausdrucksstärker sein muss.

Horrorfilm:

Ein Horrorfilm besitzt eine erhöhte Anzahl von Ereignissen, die nicht auf dem Bildschirm zu sehen sind, der sogenannte Off-Screen-Raum. Außerdem ist die Fokussierung gelegentlich uneindeutig, um Spannung zu vermitteln.

Bei der Beschreibung muss also darauf Acht gegeben werden, dass nicht zu viel verraten wird.

Dokumentarfilm:

In einem Dokumentarfilm sind die Kameraeinstellungen von Bedeutung. Nah- und Halbnahaufnahmen werden eingesetzt, um die Spezialisten besser fokussieren zu können. Wenn die Weite einer Landschaft aufgezeigt werden soll, werden oft lange Einstellungen verwendet.

Demnach sollten die Kameraeinstellungen in der Beschreibung erwähnt werden. Außerdem sollte eine ausführliche Recherche zur Genauigkeit der Terminologie stattfinden. Wenn in Naturdokumentationen Liebschaften und Feindschaften gezeigt werden, ist es ratsam, die Audiodeskription etwas subjektiver zu formulieren.

Romanze:

Eine Romanze verfügt oft über längere Bilder und zielt darauf ab, Emotionen beim Publikum zu wecken.

Daher ist häufig genug Zeit für die Audiodeskription und lyrische Formulierungen können angebracht sein. Jedoch sollte auf die Wahrung des Gleichgewichts zwischen subjektiver und objektiver Beschreibung geachtet werden (*ADLAB Audio Description guideline, o. J.*; Fryer, 2016).

Für Fryer (2016, S.104) ist die relevante Unterscheidung jedoch zwischen Actionfilm und Nicht-Actionfilm.

Mazur (2020) hingegen unterscheidet in der bereits erläuterten Analyse des Ausgangsmaterials zwischen fünf multimodalen Textsorten.

Diese werden in Abbildung 3 aufgezeigt.

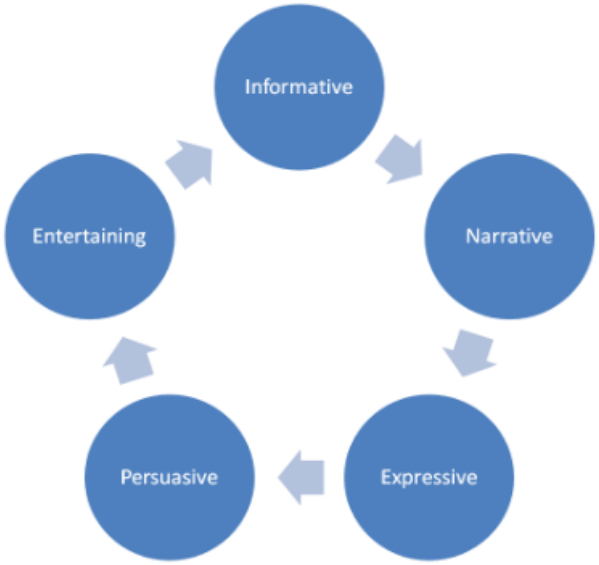

Abbildung 3: Klassifizierung multimodaler Textsorten für AD-Zwecke (Mazur, 2020, S.230)

Es existiert der informative Typ, der Dokumentationen, Nachrichten, Reise- und Kochsendungen einschließt. Dieser zeichnet sich durch eine Audiodeskription mit einer möglichst exakten Vermittlung des Inhalts aus.

Der narrative Typ, zu welchem fiktionale Filme, Seifenopern und Comicstrips zählen, versucht in der Audiodeskription möglichst die Erzählung zu rekreieren. Dies gelingt durch die Nachbildung der für die Handlung wesentlichen visuellen Elemente.

Zum expressiven Typ zählen Kunstfilme und experimentelles Kino. Die Konzentration in der Beschreibung beruht hier auf der Form und der künstlerischen Wiedergabe.

Der persuasive Texttyp beeinflusst wiederum das Verhalten und das Denken des Zuschauers, beispielsweise in der Werbebranche.

Der entertainende bzw. der unterhaltsame Texttyp umfasst Game-Shows, Wettbewerbe und Reality-Shows. Es ist wichtig, in der Audiodeskription alle unterhaltsamen Aspekte, die das Publikum zum Lachen bringen, zu erwähnen.

Diese Strategien in Bezug auf die Textsorte bestimmen die Wahl der zu beschreibenden Elemente auf der Mikroebene, um damit erneut auf die Filmanalyse von Mazur aus Kapitel 3.2 zurückzugreifen (Mazur, 2020).

3.5 Emotionstransfer durch die Audiodeskription

Um herauszufinden, auf welche Art und Weise in der Audiodeskription Emotionen vermittelt werden, ist es sinnvoll einen Blick auf den Anteil der Emotionsvermittlung zu werfen, den die Audiodeskription zum gesamten Emotionstransfer bei Filmen beiträgt.

In einer Studie untersucht Ramos (2015) das Erfahren einiger Emotionen mit und ohne Audiodeskription. Er kommt zu dem Ergebnis, dass der Emotionstransfer bei Filmen bereits zu einem großen Teil durch nicht-visuelle Komponenten vermittelt wird. Zugleich stellt er fest, dass die Emotion „Ekel" durch eine Audiodeskription deutlich verstärkt wird, da viele Informationen visuell auftreten. Weniger Audiodeskription ist zum Auslösen von „Angst" notwendig. Hierbei wird schon viel durch den Soundtrack und das Sounddesign wahrgenommen. Am wenigsten abhängig vom Sehsinn ist laut Ramos die Emotion „Traurigkeit", denn diese ist durch die Musik und durch die in der Stimme liegenden Gefühle bereits deutlich spürbar.

Anishchenko (2020, S.29) teilt den Emotionstransfer in drei Kanäle ein.

In den verbalen Kanal, der das Sprechen und Schreiben impliziert, in den paraverbalen Kanal, zu dem die Stimmlage und der Sprechrhythmus zählen und in den visuellen Kanal, welcher unter anderem die Mimik und Gestik beinhaltet.

Wie in Tabelle 1 dargestellt, vergleicht Anishchenko die zur Verfügung stehenden Mittel der klassischen Face-To-Face Kommunikation mit denen der Audiodeskription.

Tabelle 1: Vergleich zwischen dem Emotionstransfer in der Face-to-Face-Kommunikation und in der Audiodeskription (in Anlehnung an Anishchenko, 2020, S.26)

Face-to-face-Kommunikation		Audiodeskription	
verbal	• Sprechen • Schreiben • Gebärdensprache	verbal	• Sprechen • Schreiben
paraverbal	• Lautstärke, Intonation, Stimmlage und Tonhöhe • Sprechrhythmus und Sprechgeschwindigkeit	paraverbal	• Lautstärke, Intonation, Stimmlage und Tonhöhe • Sprechrhythmus und Sprechgeschwindigkeit
visuell	• Mimik • Gestik • Pantomimik • Körperliche Zustände	visuell	X

Dabei wird deutlich, dass in der Audiodeskription der fehlende visuelle Kanal durch den verbalen und paraverbalen Kanal kompensiert werden muss. Die Mimik, Gestik, Pantomimik und die körperlichen Zustände werden also sprachlich beschrieben. Es wird jedoch stark diskutiert, wie subjektiv bzw. objektiv diese Beschreibungen sein dürfen. Auf der einen Seite stellt das Interpretieren des kinetischen Verhaltens eine effektive Verkürzung der Beschreibung dar, auf die Nichtsehende angewiesen sind. Auf der anderen Seite, birgt eine subjektive Beschreibung die Gefahr zu einer Rezeptionslenkung, wodurch der Film als Kunstwerk verändert wird (Anishchenko, 2020; Poethe, 2005, S.46; Yos, 2005, S.101). Es stellt sich die Frage, ob Formulierungen wie „sie blickt sich hektisch um" oder „eine apathische Frau" bereits zu stark gewertet sind (Poethe, 2005, S.46). Offensichtlich ist, dass ein Kompromiss auf dem schmalen Grat zwischen objektiver und subjektiver Beschreibung gefunden werden muss. Benecke bemängelt, dass jegliche Form der Emotionsbeschreibung als Versuch der Überinterpretation des Filmgeschehens und der Manipulation der Emotionen sehbehinderter Rezipienten, wahrgenommen wird (Benecke, 2014). Fryer (2016, S.104) hingegen berichtet von einem Nutzer, der viele Beschreibungen, wie beispielsweise „Jenny an attractive blond" durchaus als störend empfindet.

Um ein Meinungsbild von Nutzenden bezüglich dieses Problems zu erhalten, haben Chmiel & Mazur (2022) eine Studie durchgeführt, in der 50 sehbeeinträchtigte oder blinde Menschen nach ihrer Präferenz zu gewerteten und objektiven Formulierungen befragt wurden.

In Abbildung 4 sind drei Beschreibungen aufgezeigt, wobei die fettgedruckte Version, der gewerteten Formulierung entspricht. Die erste Prozentzahl der Präferenzen bezieht sich auf die gewertete Formulierung, während die zweite Prozentzahl die Präferenz der objektiven Beschreibung angibt.

Descriptions	Preference	Chi-square test	*p* value (marked with * if significant)
an attractive singer vs. a long-legged singer in a miniskirt	43% 57%	X^2=1.82	*p*>.05
a house with peeling plaster vs. **a derelict house**	48% 52%	X^2=0.80	*p*>.05
autumn sun vs. weak sunlight	49% 51%	X^2=0.58	*p*>.05

Abbildung 4: Präferenzen zu gewerteter und objektiver Beschreibung einer AD (Chmiel & Mazur, 2022)

Die objektive Formulierung „eine langbeinige Sängerin in Minirock" wird im Vergleich zur subjektiven Formulierung „eine attraktive Sängerin" bevorzugt. Nicht so eindeutig, aber dennoch die Mehrheit präferiert auch bei den zwei Beispielen „ein Haus mit abblätterndem Putz"

im Vergleich zu „ein verwahrlostes Haus“ und „schwaches Sonnenlicht“ verglichen mit „Herbstsonne“ die erstere, objektive Formulierung.

Die hier festzustellende Tendenz der Bevorzugung von objektiven Formulierungen wird in der Forschung von Walczak und Fryer (2017) widerlegt. In deren Studie wird die gewertete Formulierung als „kreative“ Beschreibung bezeichnet und ist durch subjektive Beschreibungen der Figuren und die Einbeziehung von Elementen der Kameraführung definiert. Die Reaktionen auf die kreative Audiobeschreibung sind positiv und werden durch das intensive Seherlebnis und das starke Präsenzgefühl begründet. Jedoch herrscht ein signifikanter negativer Zusammenhang zwischen der Präferenz für die kreative Audiodeskription und dem Grad der Verwirrung der Befragten. Interessanterweise gab es keine Relation zwischen Genuss und Verständnis (Walczak & Fryer, 2017).

Es lässt sich daher schlussfolgern, das die objektive Audiodeskription vor allem dem Verständnis dient, während die kreative, subjektive Audiobeschreibung die Emotionsvermittlung stärkt und das Seherlebnis intensiviert.

Ein Lösungsansatz der Problematik zwischen objektiver und subjektiver Beschreibung ist die Berücksichtigung des Filmgenres, so wie es in Kapitel 3.4 erläutert wurde. Außerdem kann das Schreiben in Teams durch die Einschätzung mehrerer Personen hilfreich sein.

Ein weiterer Aspekt der Beeinflussung des Emotionstransfers ist die Stimme, die den Inhalt der Audiodeskription vermittelt. Mit einer sachlichen Sprechweise, die klar, deutlich und dialektfrei ist, hebt sich der Sprecher oder die Sprecherin von den Dialogstimmen ab. Dies ist in den allgemeinen Richtlinien vorgesehen und beispielsweise in der Audiodeskription der meisten Tatortfolgen umgesetzt (*ADLAB Audio Description guideline*, o. J.). Etwas emotional involvierter kann die Sprechweise bei komischen Filmen sein, jedoch nur in Maßen (Fix, 2005, S.46; Jüngst, 2020, S.118).

Nach wir vor ist jedoch die Hauptaufgabe der Audiodeskription die Verständnisvermittlung und nicht die Emotionsvermittlung.

4 Sprachsynthese

Ein Team benötigt für 90 min Film laut Fix (2005, S.45) ungefähr 5-7 Tage für die Produktion der Audiodeskription. Somit betragen die Kosten pro Film ungefähr 5000 Euro.

Um die Kosten für die Produktion von Audiodeskriptionen zu senken und somit möglicherweise die Zahl von Videos und Filmen mit Audiodeskriptionen zu steigern, werden neue Ansätze entwickelt.

Einer davon beruht darauf, die Sprachsynthese in die Produktion einzubeziehen und somit die Kosten für das Tonstudio und die Sprechenden einzusparen.

4.1 TTS-Synthese

In der Sprachsynthese wird zwischen der TTS-Synthese (text-to-speech synthesis) und der CTS-Synthese (concept-to-speech synthesis) unterschieden. Ersterem liegt ein Text als Ausgangsformat zu Grunde, wohingegen bei der CTS-Synthese die auszugebende Meldung selbstständig generiert wird. Diese automatische Generierung wird in Kapitel 5.1 erläutert. Im aktuellen Kapitel wird das System der TTS-Synthese genauer erklärt.

Um jeden beliebigen Text in ein Sprachsignal umwandeln zu können, muss dieser erst wie in Abbildung 5 dargestellt, durch die Transkription in eine Lautschrift übersetzt werden. Die Lautschrift wird als phonologische Darstellung bezeichnet und beinhaltet Informationen zu den Lauten, der Akzentuierung und der Phrasierung. Mit diesen Informationen kann schließlich in der phonoakustischen Stufe ein Sprachsignal synthetisiert werden (Pfister & Kaufmann, 2017, S.27).

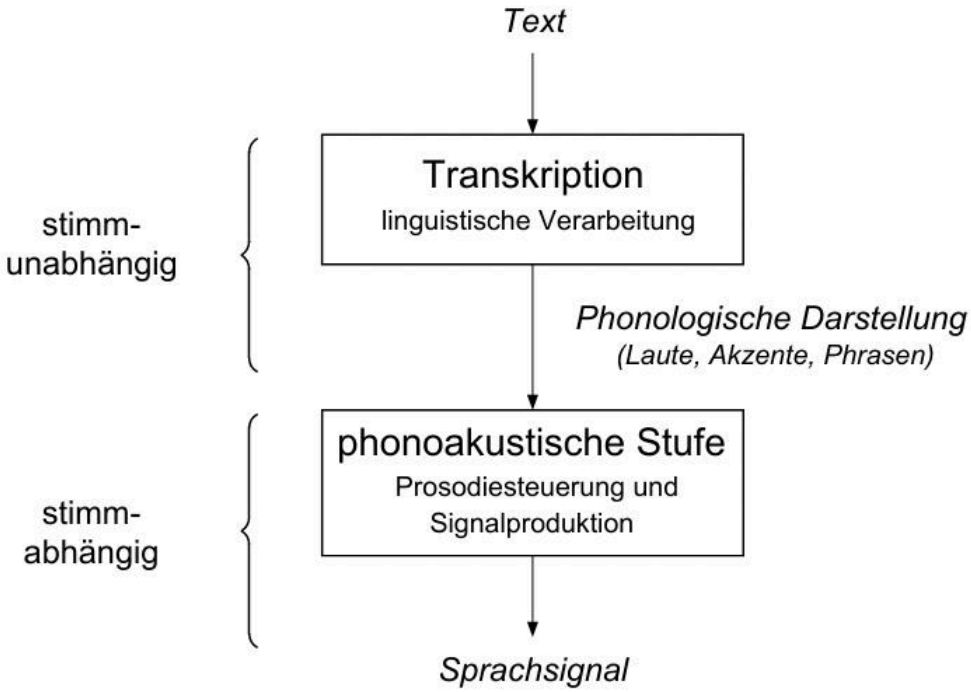

Abbildung 5: Komponenten der TTS-Synthese (Pfister & Kaufmann, 2017, S.197)

Die Vorgänge der Transkription und der phonoakustischen Stufe werden in den folgenden Unterkapiteln genauer beschrieben.

4.1.1 Transkription

Die Übersetzung des orthografischen Texts in die phonologische Darstellung erfolgt über eine linguistische Textverarbeitung, welche aus einigen Analysen besteht.

Das Ziel dieser Analysen ist es, aus den diskreten, klar voneinander getrennten Zeichen der geschriebenen Sprache eine Anweisung für die kontinuierliche, von Nachbarlauten und der Stellung im Satz abhängigen gesprochenen Sprache zu erstellen. Dies gelingt durch die *morphologische Analyse*, die *syntaktische Analyse* und die *semantische* bzw. *kontextuelle Analyse*, aus deren Ergebnissen die Akzentuierung und die Phrasierung abgeleitet werden. Zudem bedarf es anfangs einer Entschlüsselung der Abkürzungen und Akronyme des Textes (Fellbaum, 2012, S.348ff.). Abbildung 6 dient der Veranschaulichung dieses Vorgangs.

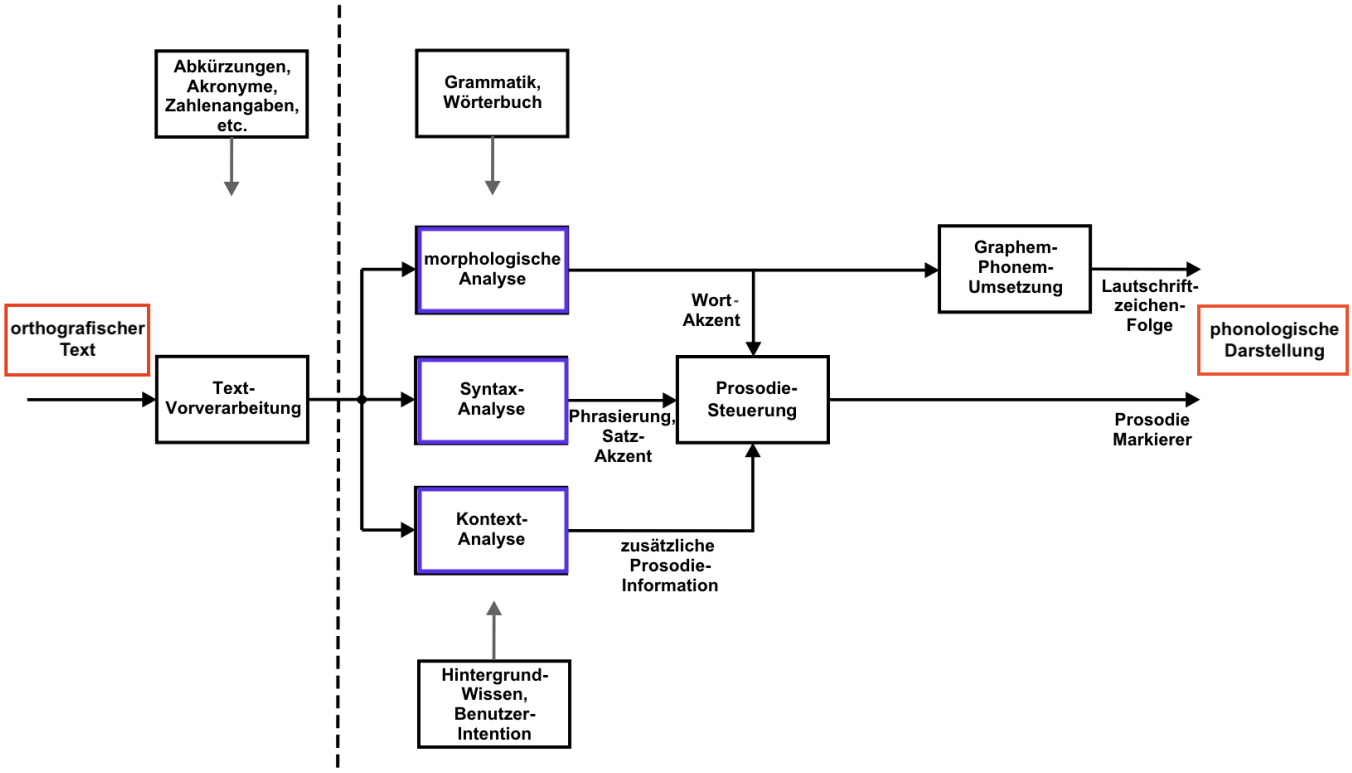

Abbildung 6: Transkription: Übersetzung des Textes in die phonologische Darstellung (in Anlehnung an Fellbaum, 2012, S.350)

Die *morphologische Analyse* befasst sich mit der Wortstruktur und ihrer Funktion. Durch das Zerlegen der Wörter in Morpheme, die kleinsten bedeutungstragenden Einheiten einer Sprache, werden alle möglichen Bedeutungsmöglichkeiten eines Wortes aufgezeigt. Die richtige Möglichkeit auszuwählen ist jedoch nicht immer einfach, denn es gibt Fälle, bei denen die korrekte morphologische Zerlegung nicht eindeutig ist. Das Wort „Wachstube“ lässt sich beispielsweise in „Wach“ + „stube“ oder in „Wachs“ + „tube“ zerlegen. Für eine Entscheidung bedarf es einer Heranziehung der syntaktischen und semantischen Analyse (Fellbaum, 2012, S.352).

Die *syntaktische Analyse* ermittelt die Satzstruktur und untersucht, ob Wortformen zu Satzteilen und diese wiederum zu Sätzen kombiniert werden dürfen. Die syntaktische Struktur trägt außerdem zur Ermittlung der Phrasierung bei, denn je stärker Wörter in der Syntax voneinander getrennt sind, desto plausibler ist es, dass sich dort eine Phrasierungsgrenze befindet. Um auf die Akzentuierung zu schließen ist die syntaktische Analyse häufig nicht ausreichend und eine semantische Analyse muss herangezogen werden. Ein Beispiel ist der Satz: „Hans beobachtet den Mann mit dem Teleskop“ (Pfister & Kaufmann, 2017, S.223). Es ist nicht klar, ob der beobachtete Mann ein Teleskop besitzt oder er durch Hans mit dem Teleskop beobachtet wird.

Die *semantische* bzw. *kontextuelle Analyse* untersucht die Bedeutung und somit die Sinnhaftigkeit von Wörtern, Ausdrücken und Sätzen. Auf der Grundlage dieser Ergebnisse wird die Prosodie, also die Akzentuierung und Phrasierung, gesteuert. Es wird beispielsweise entschieden, ob „Heroin“ wie eine Heldin oder eine Droge betont wird und ob „modern“ etwas Zeitgemäßes oder etwas Fauliges beschreibt (Fellbaum, 2012 S.353).

Zur Erstellung der Lautsprache gibt es für jede Sprache ein Lautinventar, das die Aussprache beliebiger Wörter dieser Sprache symbolisch beschreibt. Auch Phrasen, Akzente und Silbengrenzen sind in dieser symbolischen Sprache enthalten.

Im Folgenden Beispiel wird die phonologische Darstellung des Wortes „Heinrich“ aufgezeigt. Es wird das IPA-Lautschriftzeichen-System verwendet, welches das meistgenutzte System darstellt (Pfister & Kaufmann, 2017, S.198).

(P) [1]haɪn-rɪç #{2}

Abbildung 7: phonologische Darstellung von "Heinrich" (Pfister & Kaufmann, 2017, S.198)

(P)	Phrasentyp; „P“ steht für progredient, d.h. die Phrase steht nicht am Satzende
[1]	Akzentstärke der nachfolgenden Silbe; „1“ bezeichnet die stärkste Betonung
#{2}	bezeichnet eine Phrasengrenze; „1“ bezeichnet die stärkste Trennung
-	Silbengrenze

4.1.2 Phonoakustische Stufe

Um nun aus der phonologischen Darstellung ein konkretes Sprachsignal zu erzeugen, gibt es drei Hauptansätze, die schon seit einiger Zeit bekannt sind.

1939 hat Charles Wheatstone bereits einen Apparat gebaut, der den menschlichen Vokaltrakt nachahmt. Daraus entwickelte sich in den 50er Jahren der *artikulatorische Ansatz*. Er wird jedoch nicht verwendet, da die Qualität große Mängel aufweist und der Aufwand hoch ist.

Der zweite Ansatz besteht darin, das Signal zu modellieren. Das bekannteste Modell hierfür ist die *Formant-Synthese*, welche ebenfalls zu den ältesten Syntheseverfahren zählt. In elektronischen Schaltkreisen werden die Resonanzen des Vokaltrakts mithilfe von Filtern simuliert. Es wird zwischen stimmhafter Sprache, welche ein quasiperiodisches Signal erzeugt, und stimmloser Sprache, welches ein rauschartiges, aperiodisches Signal vorweist, unterschieden. Bis Mitte der 1980er Jahre war dies das Standardmodell. Heute wird es nur eingesetzt, wenn wenig

Speicherplatz und eine begrenzte Rechenleistung zur Verfügung steht (Gold et al., 2011, S.432f.).

Die qualitativ besten Resultate werden mit dem heutzutage am häufigsten verwendeten Ansatz der *konkatenativen Synthese,* oder auch *Verkettungsansatz* genannt, erzielt. Hierbei werden voraufgezeichnete natürliche Sprachsignale in Segmente unterteilt und anschließend so verkettet, dass jede beliebige Aussage möglich ist. Der Vorteil dieser Methode ist, dass es sich um absolut natürliche Sprache handelt. Als erster Schritt dieser Synthese muss ein Inventar der zu verkettenden Segmente angelegt werden. Dieser sogenannte Korpus besteht aus Polyphonen und Diphonen, welche auch als Grundelemente bezeichnet werden. Polyphone können sich über mehrere Laute erstrecken, wohingegen Diphone lediglich zwei Laute beinhalten. Geschnitten werden die Grundelemente in der Lautmitte, wo der Laut am ausgeprägtesten ist. Nur bei Plosiven wird kurz vor der Plosion geschnitten, da dort eine Pause entsteht. Der Übergang von einem Grundelement ins nächste wird als Stoßstelle bezeichnet. Dort sollten möglichst geringe Diskontinuitäten herrschen. Um die Anzahl der Stoßstellen zu verringern und somit eine gute Prosodie zu erzeugen, ist es sinnvoll möglichst lange Polyphonsegmente zu verwenden. Diese Methode wird Korpussynthese genannt. Allerdings wird hierbei eine große, variantenreichen Sammlung benötigt, welche eine große Menge von aufgezeichneten Sprachsignalen benötigt. Es kann daher hilfreich sein, ein anwendungsspezifisches Inventar zu erstellen, welches nützliche Wortfolgen für die gegebene Anwendung bereithält. Das Gegenteil der Korpussynthese ist die Diphonsynthese. Hier existiert von jedem Grundelement nur ein Exemplar. Somit ist das Inventar zwar kompakt, jedoch muss die Prosodie stark angepasst werden, da die Laute je nach Kontext unterschiedlich betont werden. Diese Modifizierung führt zu Qualitätsverlust (Pfister & Kaufmann, 2017).

Das gängigste Verfahren zum Anpassen der Prosodie und zur Verkettung der Sprachsignalsegmente ist das PSOLA-Verfahren. Es beinhaltet eine Veränderung der Grundfrequenz, der Dauer und der Lautstärke des Sprachsignals. Vereinfacht erklärt wird zwischen stimmhaften und stimmlosen Partien unterschieden, wobei stimmhafte Partien am Anfang jeder Periode markiert werden und stimmlose hingegen in Intervalle fester Länge unterteilt werden. Durch die Multiplikation mit einer Fensterfunktion entstehen aus zwei benachbarten Abschnitten sogenannte Doppelperiodensegmente. Durch die Vergrößerung oder Verkleinerung der Überlappung der Doppelperiodensegmente verändert sich die Tonhöhe und infolgedessen auch die Dauer. Ersichtlich wird dieser Vorgang in der zweiten Zeile der Abbildung 8. Wenn dieser Effekt nicht erwünscht ist, muss er kompensiert werden. Das erfolgt durch das Aneinanderreihen oder das Weglassen gewisser Doppelsegmente, welches in Zeile 3 und 4 der Abbildung 8 dargestellt ist (Pfister & Kaufmann, 2017, S.260ff.).

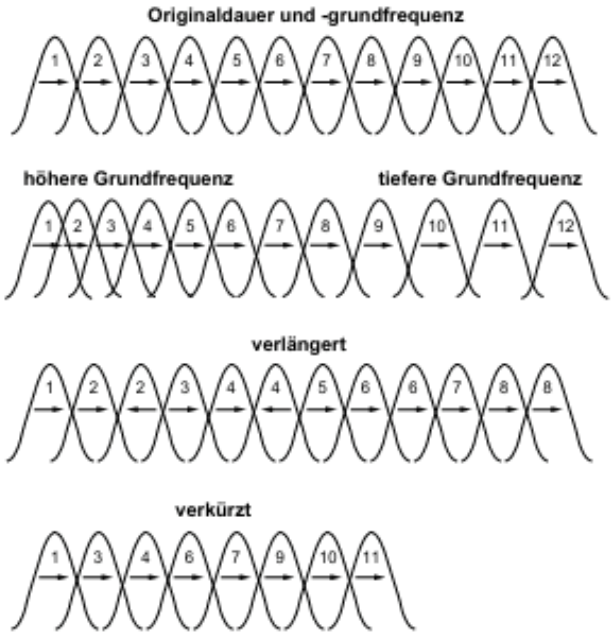

Abbildung 8: Doppelperiodensegmente als Hanning-Fenster dargestellt (Pfister & Kaufmann, 2017, S.262)

Die Grundfrequenz ist im Gegensatz zu anderen Sprachen in der deutschen Sprache nicht bedeutungsentscheidend, hat jedoch die Aufgabe der prosodischen Steuerung. Außerdem sinkt sie während des Sprechens leicht ab, was durch den abfallenden Druck in der Luftröhre verursacht wird. Dieses Phänomen wird Deklination genannt. Die Grundfrequenz kann aber auch bei gewissen Phrasen, wie zum Beispiel am Ende eines Fragesatzes, ansteigen.

In folgender Abbildung 9 ist der Grundfrequenzverlauf und das zugehörige Sprachsignal des Satzes „Sie erhielten bei ihrem Zug durch die Straßen Zulauf von Pekinger Bürgern“ abgebildet. Der obere Teil der Abbildung stellt die zugehörige Sprachaufnahme dar.

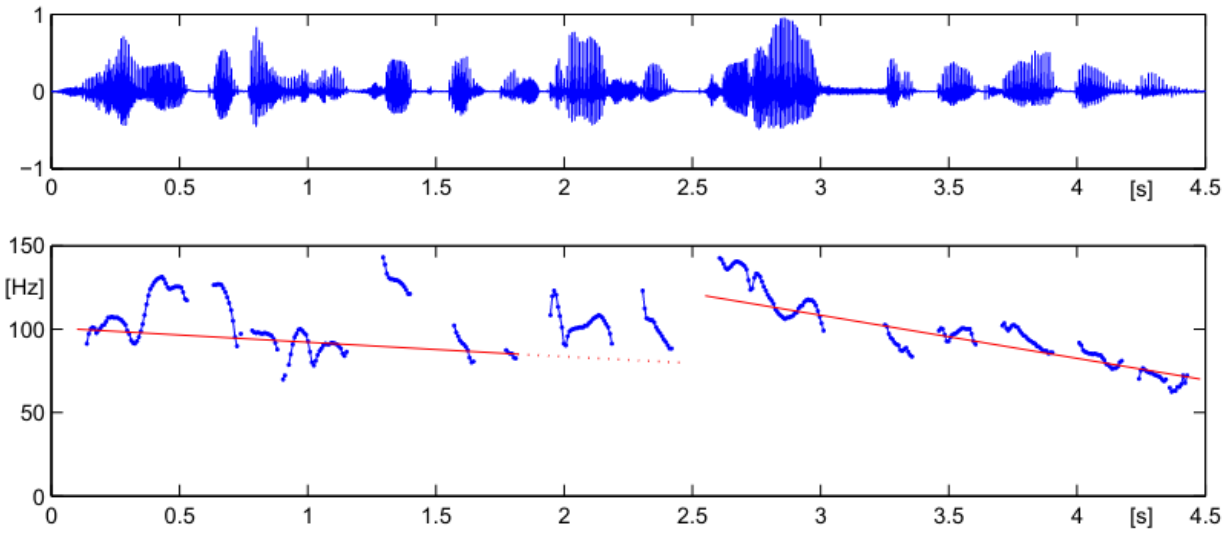

Abbildung 9: Grundfrequenzverlauf eines Satzes mit zwei Phrasen (Pfister & Kaufmann, 2017, S.274)

Es kann festgestellt werden, dass die erste Phrase am Ende etwas ansteigt, deshalb ist die Linie an dieser Stelle gepunktet abgebildet. Die fallende Deklination im Verlauf einer Phrase lässt sich jedoch klar erkennen. Um den Grundfrequenzverlauf eines Satzes herauszufinden, müssen die Deklinationsgrade bestimmt werden. Diese können entweder mittels eines linearen Ansatzes oder eines neuronalen Netzes berechnet werden. Die Resultate durch neuronale Netze sind kaum von der natürlichen Sprache zu unterscheiden, wohingegen die erstellten Verläufe mittels linearen Ansatzes bei längeren Texten schnell eintönig wirken.

Ebenfalls wichtig zu erwähnen ist, dass sich die Lautheitswahrnehmung des menschlichen Gehörs von dem tatsächlichen Pegel des Signals unterscheidet, sodass benachbarte Laute als gleichlaut empfunden werden, obwohl beispielsweise das „a“ im Vergleich zum darauffolgenden „s“ einen deutlich höheren Pegel aufweist. Da es schwierig ist, eine Intensitätssteuerung zu erschaffen, die eine merkliche Verbesserung bewirkt, wird sie bei der Sprachsynthese mit dem konkatenativen Ansatz im Normalfall nicht eingesetzt.

Solche und weitere Probleme führen dazu, dass die Qualität der Sprachausgabe eines Computers noch nicht der eines Menschen entspricht. Die konkreten Schwierigkeiten werden im nächsten Kapitel aufgezeigt (Pfister & Kaufmann, 2017, S.273ff.).

4.2 Schwierigkeiten bei der Sprachsynthese

Lautsprache von einem Computer produzieren zu lassen schien in den 60er Jahren im Hinblick auf die rasante Entwicklung der Computertechnik, ein schnell zu lösendes Problem zu sein. Jedoch sind viele Forscher heute der Ansicht, dass auch in 20 Jahren noch nicht die maschinelle Sprachfähigkeit eines Menschen erreicht sein wird. Der Wortschatz wird zwar weitestgehend von der Maschine beherrscht, jedoch liegt das Problem bei der Einschätzung der richtigen der Syntax und bei der Bedeutung von Sätzen, die sich in Bezug auf den Kontext stark ändern können. Ein Beispiel für eine fehlerhafte syntaktische Analyse wurde bereits im vorherigen Kapitel 4.1.1 gegeben, bei dem der Satz „Hans beobachtet den Mann mit dem Teleskop" als Beispiel eines schwierig zu analysierenden Satzes gegeben wurde. Es gibt jedoch auch Sätze, die sich aufgrund ihrer unvollständigen Syntax gar nicht erst zuordnen lassen. Ein Beispiel hierfür ist: „Nichts wie weg!". In diesem Fall muss ein Ersatzsyntaxbaum erstellt werden, bei dessen Generierung oft Fehler entstehen. Die semantische Analyse, also die Bedeutung von Sätzen, lässt sich für ein künstliches System oft schlicht nicht korrekt durchführen, da das nötige Weltwissen fehlt und das Berücksichtigen von Beziehungsebenen zu anspruchsvoll für eine Maschine ist (Braun & Starr, 2022, S.397; Pfister & Kaufmann, 2017, S.21f.). Sprachassistenten wie Alexa oder die Google-Assistentin verstehen bisher beispielsweise nur die Syntax. Die Semantik spielt jedoch ebenfalls eine große Rolle. Hierzu gehört auch die Erkennung von Humor, Ironie und Emotionen (Buchkremer, 2020).

Laut Gold et al. (2011) ist die Tatsache, dass es mehr als ein akzeptables Prosodie-Muster gibt, eine weitere große Herausforderung.

Neben diesen Hauptproblemen gibt es weitere Schwierigkeiten wie nicht analysierbare Wörter durch fehlende Lexikoneinträge. Hierzu zählen Eigennamen oder Fremdwörter, die Aussprachefehler verursachen können. In solchen Fällen wird aus dem Graphem-zu-Phonem Prinzip, also aus der orthografischen Schreibweise, eine direkte Aussprache mit den Regeln der eingestellten Sprache generiert (Pfister & Kaufmann, 2017).

Nicht zuletzt ist eine gewisse Eintönigkeit der künstlichen Stimme nicht zu vermeiden, da sie anhand vorgegebener Regeln produziert wird (Gold et al., 2011).

4.3 Menschliche Stimme vs. synthetische Stimme

Die Bereitstellung von barrierefreien Angeboten wie die Audiodeskription erfordern zum Teil einige kosten- und zeitintensive Produktionsschritte.

Technologische Werkzeuge, wie die Nutzung einer synthetisierten Stimme, können diese Prozesse verkürzen. Allerdings sind solche Veränderungen oft stark umstritten und bringen sowohl Vorteile als auch Nachteile mit sich.

Zuerst sollte überprüft werden, ob tatsächlich Kosteneinsparungen bei den gegebenen Bedingungen erreicht werden können, denn Nakajima & Mitobe (2022) haben festgestellt, dass die Anpassungsarbeiten beim Einsatz von künstlichen Stimmen ebenfalls viel Zeit in Anspruch nehmen können. Vor allem die Korrektur von Einfüge-Punkten kann bei der Verwendung von einfachen Methoden wie Excel-Tabellen zeitaufwändig sein. Daher muss erst ein passender Workflow und geeignete Software für die Erstellung einer Audiodeskription mit künstlicher Stimme gefunden werden. Wenn diese Bedingung erfüllt ist, kann die Flexibilität einer künstlichen Stimme von großem Nutzen sein. Verschiedene Sprachen, Geschlechter und Stimmfarben können mit wenigen Mausklicken eingestellt werden (Kurch, 2019, S.444).

Außerdem würde der Bedarf an menschlichen Sprecherinnen und Sprechern nicht zwingend geringer werden, sondern viel mehr würde der Einsatz von künstlichen Stimmen die Anzahl von audiobeschriebenen Filmen erhöhen (Kurch, 2019, S.45; Szarkowska, 2011). Unter diesem Aspekt wird in mehreren Studien eine hohe Akzeptanz von künstlichen Stimmen bestätigt. In der von Matamala & Orero (2016, S.280) beschriebenen Studie wurde die natürliche Stimme statistisch besser bewertet als die Künstliche, jedoch fanden 94% der Teilnehmenden die synthetisierte Audiodeskription eine akzeptable Alternative. 20 % der Testpersonen haben die künstliche Stimme bevorzugt. Die Studie von Kobayashi et al. (Rocha Façanha et al., 2016, S.507), sowie die Ergebnisse der Studie von Szarkowska (2011, S.154) unterstützen diese Aussage. In Abbildung 10 sind die Ergebnisse der zuletzt genannten Studie zur Frage, ob sich die 24 Testpersonen die künstliche Stimme als Übergangslösung oder als Dauerlösung vorstellen können, graphisch dargestellt.

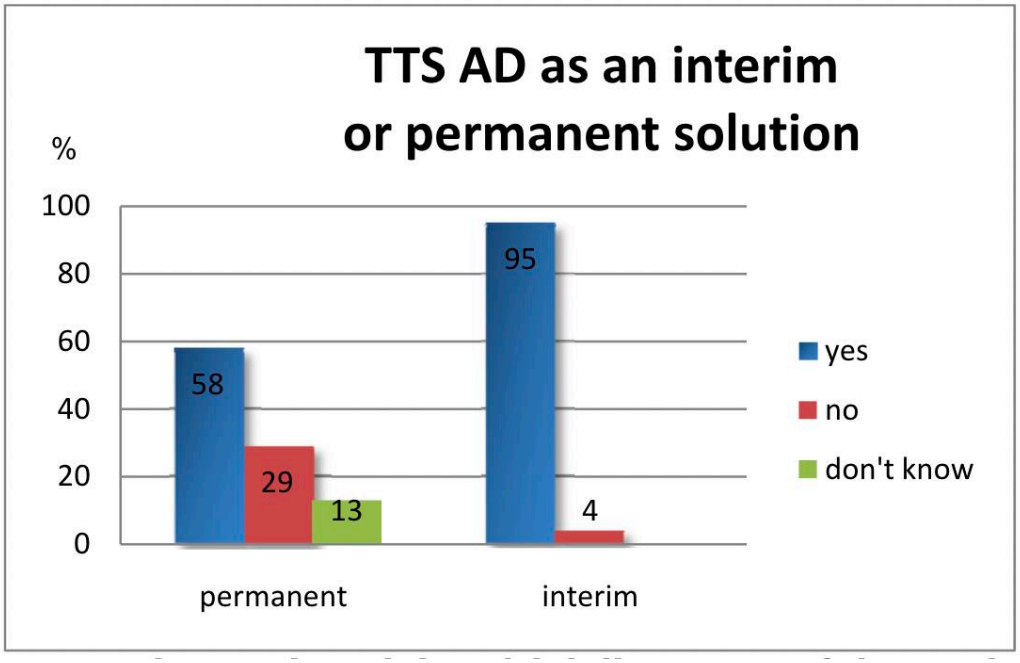

Abbildung 10: Akzeptanz der TTS-AD als Dauerlösung oder Übergangslösung (Szarkowska, 2011, S.154)

„Die Stimme sei weitaus weniger roboterartig als man denkt“, kommentiert Netzwoche-Redaktor René Jaun (2021) nach dem Hören einer künstlich vertonten Audiodeskription. Jedoch bemängelt er die Aussprache beim Einfluss von Fremdsprachen (Jaun, 2021). Die qualitativ unterlegene synthetisierte Stimme weist ebenfalls Defizite in der Emotionsvermittlung auf. Fryer und Freeman haben 2014 in einer Studie festgestellt, dass nur mit menschlicher Stimme ein höheres Maß an Präsenz hervorgerufen und die Emotionserweckung verbessert werden kann. Es ist jedoch zu ergänzen, dass bei nichtemotionalen Inhalten kein statistisch signifikanter Unterschied bei der Emotionserzeugung zwischen den Stimmen ermittelt wurde (Walczak & Fryer, 2017). Die Erwägung der Absichten der Audiodeskription und das Einbeziehen der Eigenschaften einer Filmsorte sind daher sinnvoll. Da eine Audiodeskription objektiv und neutral klingen soll und primär eine Verständnisvermittlung beabsichtigt, ist der Standpunkt, den möglichen Emotionsverlust durch eine künstliche Stimme zu vernachlässigen, durchaus nachvollziehbar.

In der japanischen Studie von Kobayashi et al. wurde herausgefunden, dass künstliche Stimmen vor allem in dokumentarischem Filmmaterial und Lehrvideos nützlich sein können, wohingegen bei unterhaltendem Filmmaterial eine menschliche Stimme von den Konsumierenden bevorzugt wird (Campos, 2020; Walczak, 2018). Die gleiche Erkenntnis wird auch auf der Webseite der ADLAB-Guidelines aufgeführt (*ADLAB Audio Description guideline*, o. J.).

Eigene Untersuchungsergebnisse bezüglich dieses Themas werden im praktischen Teil der Arbeit (Kapitel 6) hervorgebracht.

5 Stand der Technik

In diesem Kapitel wird ein Überblick über die neuesten technologischen Fortschritte, Trends und Herausforderungen im Bereich der Audiodeskription vermittelt. Dadurch werden potenzielle Zukunftsperspektiven aufgezeigt und das Grundlagenwissen für die Entwicklung des in Kapitel 6 beschriebenen Prototyps geschaffen.

5.1 Automatische Generierung des AD-Skripts

Die Produktion einer Audiodeskription lässt sich grundlegend in zwei Schritte einteilen. Der erste Schritt beinhaltet die Erstellung eines Beschreibungsskripts, wohingegen sich der zweite Schritt mit der Erzeugung der gesprochenen Sprache befasst. Letzterer Schritt basiert auf der Technologie der TTS-Synthese, welche einen fortgeschrittenen technologischen Stand vorweist und in Kapitel 4 ausführlich beschrieben wurde.

Wie Expertinnen und Experten feststellen mussten, gestaltet sich eine Automatisierung im Bereich der Skript-Erstellung schwieriger. Die Versuche reichen weit in die Vergangenheit, denn schon 2002 wurde an der Universität Surrey versucht, auf der Grundlage von Regieanweisungen ein AD-Skript automatisch herzustellen. Die Ergebnisse waren gut, wenn auch mit Abstrichen (Jüngst, 2020, S.109).

Seit einigen Jahren wird nun auch mit Machine-Learning zur Erstellung der Skripte experimentiert.

Eine naheliegende Methode ist das Einspeisen von bereits erstellten Audiodeskriptionen in ein System, das aus diesem Material lernt. Das Trainierte wird dann auf neues audiovisuelles Material angewendet. Dabei können die Transkription und die Zeitkodierung inklusive Metadaten automatisch generiert und anschließend beispielsweise in einer SRT-Datei gespeichert werden. Da dies zum jetzigen Zeitpunkt nicht fehlerfrei möglich ist, bedarf es einer menschlichen Nachbearbeitung (Kurch, 2019, S.447ff.).

Ein fortschrittlicherer, von Campos et al. (2020, S.101) beschriebener Ansatz zur technischen Erstellung der Beschreibung beruht auf einer Kombination aus der Analysierung der visuellen Informationen mittels Objekterkennungsverfahren und der Analysierung der Informationen im Drehbuch. Für die Objekterkennung werden CNNs (convolutional neural networks) und RNNs (recurrent neuronal networks) eingesetzt. CNNs, übersetzt „gefaltete neuronale Netzwerke“, erkennen typische Merkmale von Objekten durch Training und identifizieren sie. Auch Beziehungen zwischen Objekten können festgestellt werden. Die RNNs, übersetzt „rekurrente neuronale Netze“, dienen zur Verarbeitung und Interpretation sequenzieller Informationen

und geben mit Wahrscheinlichkeit bewertete Phrasen oder Sätze zur Beschreibung zurück. Bei der Analysierung des Drehbuchs werden relevante Informationen anhand der am häufigsten verwendeten Wörter extrahiert und diese in Sätze konvertiert. Das Problem hierbei ist, dass keine Zeitinformationen verfügbar sind, die den Text mit dem Video verknüpfen. Daher muss weitere Software eingesetzt werden, um die Stillen und somit die Lücken für die Audiodeskription zu ermitteln. Der Einsatz von Objekterkennung und einer Drehbuchanalyse ist sehr komplex, sodass das Ergebnis letztendlich nicht zufriedenstellend ist. Das mit dieser Methode entwickelte CineAD-System wird in Kapitel 5.3 vorgestellt und die dazugehörige Studie beschrieben.

Braun & Starr (2019) kritisieren, dass das Trainieren einer KI durch bereits vorhandene, professionell erstellte Audiodeskription deshalb keine Qualität bietet, weil die bereits durch die Dialoge und Geräusche vermittelten Informationen außen vorgelassen werden. Diese spielen eine große Rolle, sind aber filmspezifisch unterschiedlich. Bei der Objekterkennung argumentieren sie, dass zwar große Fortschritte im Bereich der Gesichtserkennung und der Erkennung von Emotionen erreicht wurden, jedoch die Übertragung auf narrative audiovisuelle Inhalte weiterhin eine große Herausforderung darstellt. In den neuesten Forschungen von Braun & Starr (2022) werden aus diesen Gründen Trainingssätze aus natürlichsprachlichen Beschreibungen für die Erkennung und Beschriftung visueller Objekte verwendet. Große Datensätze wie MS COCO enthalten menschliche Beschreibungen zu Bildern, GIFS oder kurzen Videoclips. Dieser Ansatz wurde auch im H2020 Projekt MeMAD (*Methods for Managing Audiovisual Data*) verfolgt. Da eine Aneinanderreihung von Beschreibungen zu einer kohärenten, linearen Handlung erforderlich ist, die Maschine aber nur Einzelbilder bzw. kurze Sequenzen erkennt, wurden im Rahmen des MeMAD-Projekts Spielfilme in kleine, in sich geschlossene erzählerische Einheiten zerlegt und somit ein Korpus erstellt. Insgesamt wurden 501 Ausschnitte aus 44 Spielfilmen bearbeitet. Die Studie zu diesem Ansatz stellt fest, dass die menschlichen Beschreibungen kreativer und unterhaltungsorientierter als die von Maschinen erstellten Texte sind und zudem häufig Schwierigkeiten bezüglich der Relevanz der gewählten Informationen auftreten. Außerdem existieren auch bei dieser Methode grundlegende Probleme wie Fehlinterpretationen und Verwechslungen, was unter anderem an einer mangelnden Verfügbarkeit und Existenz der Trainingssätze liegt. Ein Beispiel für eine Verwechslung ist die Fehlidentifikation eines Surfbretts, das von der KI als Schreibtisch erkannt wird (Braun & Starr, 2022, S.396f.).

Die Erstellung von AD-Texten ist ein hochkomplexer Prozess der intersemiotischen Übersetzung, bei welchem zum jetzigen Zeitpunkt noch keine Methode weit genug entwickelt ist, um zufriedenstellende Ergebnisse zu liefern. Darum wird eine unterstützende Funktion anstelle einer Ersetzung der menschlichen Arbeit vorgeschlagen. Außerdem wird von Fachleuten ein

enormes Potenzial dieser Anwendung im Bereich von Social-Media-Inhalten gesehen (Braun & Starr, 2022, S.398)

5.2 Synthetische Stimmen

Auch weil künstliche Stimmen stark umstritten sind, werden sie bisher erst selten für Audiodeskriptionen eingesetzt. Beispiele, bei denen eine künstliche Stimme die AD spricht, sind die ARD Webserie „How-To-Tatort“ und einige Dokumentarfilme auf Netflix (*ARD Mediathek*, o. J.; *Netflix*, o. J.). Einen Mehrwert können künstliche Stimmen außerdem zu Testzwecken in der Herstellung von Audiodeskriptions-Skripten bieten, indem ausprobiert werden kann, ob der Text in die Lücken passt.

Die fortschrittlichste Art der künstlichen Stimmen sind neuronale Stimmen, die mittels Machine Learning von voraufgezeichnetem Sprachmaterial lernen und dadurch eine besonders hohe Qualität der synthetisierten Sprache vorweisen. Diese Stimmen heißen bei einigen Anbietern „neuronale Stimmen“, bei Google tragen sie wiederum den Namen „WaveNet-Stimmen“ und andere Firmen geben ihnen ebenfalls eigenkreierte Namen. Verfügbar sind diese Stimmen entweder als lokal installierte Sprachsynthese-Softwares oder als cloudbasierte Platform-as-a-service-Anwendung. Anpassungen zur Optimierung der Prosodie, Geschwindigkeit, Tonhöhe und einige weitere Einstellungen können über einen SSML-Code (eine eigene Markup-Sprache für solche Anwendungen) oder über eine graphischen Benutzeroberfläche getätigt werden.

Einige Anbieter wie Microsoft bieten mittlerweile den Dienst an, aus einer beliebigen Stimme eine neuronale, synthetisierte Stimme zu erstellen, sodass von einer personalisierten Stimme jeder beliebige Text vorgelesen werden kann (Microsoft, 2022). Durch den konkatenativen Ansatz der Sprachsynthese, welcher in Kapitel 4.1.2 der Sprachsynthese erläutert wurde, ist dies mit einfachen Mitteln möglich. Die gesamte linguistisch-phonetische Transkription bleibt unverändert, nur die Lautelement-Bibliothek für die gewünschte Stimme muss erstellt werden.

Eine große Menge an voraufgezeichnetem Audiomaterial des Sprechers oder der Sprecherin und eine Unterteilung in Segmente macht eine Erschaffung einer solchen Lautelement-Bibliothek möglich (Pfister & Kaufmann, 2017). Als empfohlene Trainingsmenge der Stimme wird von Microsoft ein Umfang von 20-40 Stunden mit 300-2000 Äußerungen verlangt, um eine gute Qualität der personalisierten Stimme zu erreichen (Microsoft, 2022).

Durch die Möglichkeit der Erstellung einer personalisierten Stimme bekommt das Thema Voice-Cloning eine wichtige Bedeutung. Es gibt zahlreiche Softwares wie Lyrebird, VoiceLab und ElevenLabs, die schon mit wenigen Trainings-Samplen die charakteristischen Merkmale

einer Stimme erkennen und Sätze produzieren können, die eine bestimmte Person so nie gesagt hat (Beuth, 2018).

Die Arbeit von Arik et al. (2018, S.1) zeigt die Verbesserung der synthetisierten neuronalen Stimme durch eine steigende Zahl an Trainingsproben anhand von Audiobeispielen auf *(Audio demos, o. J.)*. Trotzdem ist der Unterschied zwischen den menschlichen Audiobeispielen und den synthetisierten Beispielen deutlich zu hören.

Da die Nutzung solcher Programme jedoch einen geringen Aufwand darstellt und sie für jeden zugänglich sind, nimmt der Missbrauch im Bereich der Audio-Fakes zu und wird ernstzunehmender. So lassen beispielsweise Trolle Anfang diesen Jahres Emma Watson aus dem Buch „Mein Kampf“ vorlesen (Mark8-36, 2023).

Technische Mittel ermöglichen es momentan, diese Fakes ohne Probleme zu identifizieren (Böhm, 2023; Christian Schiffer, 2023).

5.3 Erweiterungen für Audiodeskriptionen

Um den Zugang zu Audiodeskription für viele Menschen zu ermöglichen und ein intensives Filmerlebnis zu erzeugen, existieren einige unterstützende Technologien wie Smartphone Apps, Software-Tools, haptische Elemente und diverse Funktionen bei Videoplayern.

Eine der wichtigsten Errungenschaften dieser Erweiterungen ist die Greta & Starks App, bei der Tonspuren mit Audiodeskriptionen zu entsprechenden Filmen bereitgestellt werden. Zu Beginn des Films synchronisiert sich das Smartphone mit dem Film-Ton. Ein Vorteil dieser App ist, dass sie ebenfalls in Kinos verwendet werden kann, wenn eine Audiodeskription verfügbar ist und diese vom Filmverleih lizensiert wurde. Außerdem können eigene Kopfhörer verwendet werden und es ist keine Bedienung eines fremden Geräts notwendig, da das eigene Smartphone verwendet werden kann. Andere Länder haben ebenfalls dementsprechende Apps entwickelt. In Italien wird die App MovieReading verwendet, in Holland die App EarCatch, in Spanien AudescMobile und in Schweden existiert eine App namens MovieTalk (Kurch, 2019, S.775).

Für eine erleichterte Produktion der Audiodeskriptionen können AD-Editor-Softwares genutzt werden. Es gibt einige auf dem Markt, jedoch haben sie grundlegend die gleiche Funktionsweise. Es existiert immer ein Quellfenster und ein Skriptfenster. Durch das Drücken einer Funktionstaste wird ein neues Skriptfenster mit Timecode geöffnet, in das die Beschreibung dieser Lücke eingefügt wird. Der Timecode erleichtert die Synchronisation zum späteren Zeitpunkt. Außerdem wird das gemeinsame Arbeiten mehrerer Personen an einem Projekt ermöglicht. Häufig besteht die Auswahl zwischen der Verwendung von synthetisierter Sprache und

eigenen Aufnahmen. Das Benutzen solcher Editor-Softwares ermöglicht zudem ein flexibles und einfaches Exportieren von Dateien.

Bekannte Softwares sind beispielsweise Fingertext, Stellar und die kostenfreie Software YouDescribe.

In der neuen Software AD Author kann Text-To-Speech eingesetzt werden, um einen Platzhalter für die spätere Sprachaufnahme zu bieten. Außerdem können wahlweise provisorisch eigene Aufnahmen als Platzhalter eingesprochen werden. Beim Exportieren der Dateien kann zwischen einer simplen Wavedatei bis hin zu einer vollständigen Videokompilierung, in der alle Audiospuren enthalten sind, gewählt werden. Diese unterstützenden Softwares werden aber nur teilweise von Fachleuten eingesetzt. Eher werden sie von Auszubildenden als Hilfestellung verwendet (Minutella, 2022, S.331ff.).

Um eine Verbesserung der standardisierten Audiodeskription zu erzielen, wurden weitere Anwendungen entworfen und getestet. Im Folgenden werden einige davon vorgestellt.

In Action-Filmen und Videos mit vielen schnellen Handlungen muss aufgrund der geringen Zeit für Audiodeskriptionen oft das Verständnis leiden. Deshalb wurde 2018 der AuDIVA-Player von Pantula & Kuppusamy (2019) entwickelt. Dieser ermöglicht ein Anhalten des Videos, um eine Beschreibung der Handlung vor dem Abspielen der beschriebenen Szene hinzuzufügen. Die Funktion wird durch HTML-, CSS- und JavaScript-Dateien umgesetzt. Aus den Ergebnissen des entsprechenden Usability-Tests resultiert eine Eignung für kurze Videos, nicht aber für längere Videos (Pantula & Kuppusamy, 2019).

Auch die international etablierten WCAG-Richtlinien für die barrierefreie Gestaltung von Internetangeboten definieren den Begriff „erweiterte Audiodeskription“ als das Anhalten des Videos für weitere Informationen, wenn ohne diese der Sinn des Videos verloren gehen würde (*Richtlinien für barrierefreie Webinhalte (WCAG) 2.1*, o. J.).

Ein weiterer Player, der sogenannte ADV-Player, wurde von Rocha Façanha et al. (2016) entwickelt und getestet. Er bedient sich der „Text-To-Speech“ Technologie und generiert aus einem vorliegenden Skript, welches die Beschreibung und die Ausführungszeiten enthält, eine synthetisierte Audiospur.

In folgender Abbildung 11 wird der Prozess der Erstellung dieser Audiospur aufgezeigt.

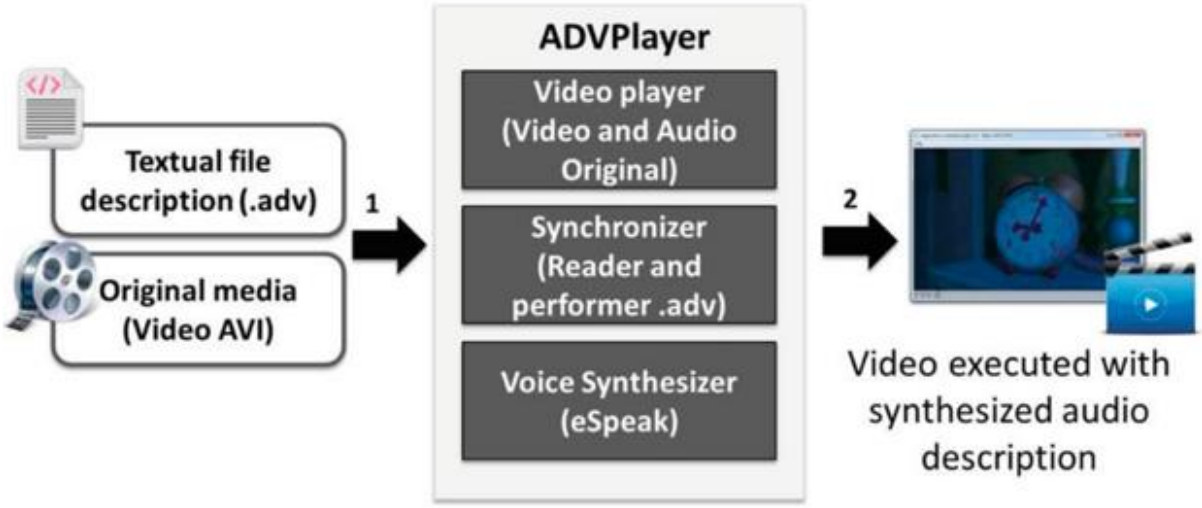

Abbildung 11: ADV-Player Komponenten (Rocha Façanha et al., 2016, S.508)

Das System basiert auf der Eingabe einer Textdatei im .adv-Format, sowie dem Hinzufügen der originalen Video-Datei. Um ein Video mit synthetisierter Audiodeskription zu erhalten, wird zuerst ein digitaler Medienplayer eingesetzt, um das Video und Audio abzuspielen. Dann kommt ein Synchronisier-Gerät zum Einsatz, das die Beschreibungen liest und den Synthesizer aktiviert. Schließlich wird der Sprachsynthesizer selbst in Form einer TTS-Software angewendet. Um die Anzahl an verfügbaren Beschreibungsskripten zu erhöhen, wurde dafür eine Webseite zum Austausch dieser Dateien errichtet.

Getestet wurde der Player an 19 Studierenden mit Sehbeeinträchtigung. Sie konsumierten 15 Minuten des jeweiligen Anfangs von drei Filmen. Als Ergebnis resultierte eine hohe Zufriedenheit und eine Akzeptanz von Audiodeskriptionen mit synthetisierter Stimme. Jedoch gab es negative Kritik bezüglich der Sprachgeschwindigkeit und Lautstärke. Diese kann bei synthetisierten Stimmen mit wenig Aufwand angepasst werden (Rocha Façanha et al., 2016). Jedoch musste auch schon Szarkowska (2011) feststellen, dass der Startpunkt der Audiodeskription bei einer Geschwindigkeitsänderung gleich ist, die Deskription aber zu unterschiedlichen Zeitpunkten endet und gegebenenfalls nicht in die Lücke passt.

In einem weiteren Projekt, wurde 2020 von Campos et al. das CineAD-System vorgestellt.

Dieses System ist ein Versuch, die Technologie der automatischen Skripterstellung mittels KI, anzuwenden. Die Schritte, die durchlaufen werden, um diese automatische Generierung umzusetzen sind folgende:

1. Lesen und Extrahieren der Elemente im Drehbuch. Dazu gehören beispielsweise Szenentitel, Handlungen, Dialoge und Figuren.
2. Lückenidentifikation anhand der Untertitel.
3. Zusammenfassung der wichtigsten Informationen des Drehbuchs durch die am häufigsten auftauchenden Wörter.
4. Erstellung des AD-Skripts mit passenden Timecodes
5. Speichern der Datei im SRT-Format

Die Sprache wird entweder mithilfe von Text-To-Speech synthetisiert oder von einer Sprecherin oder einem Sprecher eingesprochen. Der Anspruch an dieses System ist nicht, eine professionelle, qualitativ hochwertige Audiodeskription zu generieren, sondern den Sehbeeinträchtigten lediglich einen Zugang zu mehr Informationen zu verschaffen. Daher wurde das System in der Studie nicht mit einem manuell erstellten Audiobeschreibungs-Skript verglichen, sondern ausschließlich der Mehrwert des Systems im Vergleich zum Video ohne Audiodeskription getestet. Hinsichtlich dieses Aspekts resultierte eine deutlich erhöhte Verständnisvermittlung, wie in Abbildung 12 erkenntlich ist (Campos et al., 2020).

Der Test wurde mit 12 sehbeeinträchtigen Menschen durchgeführt.

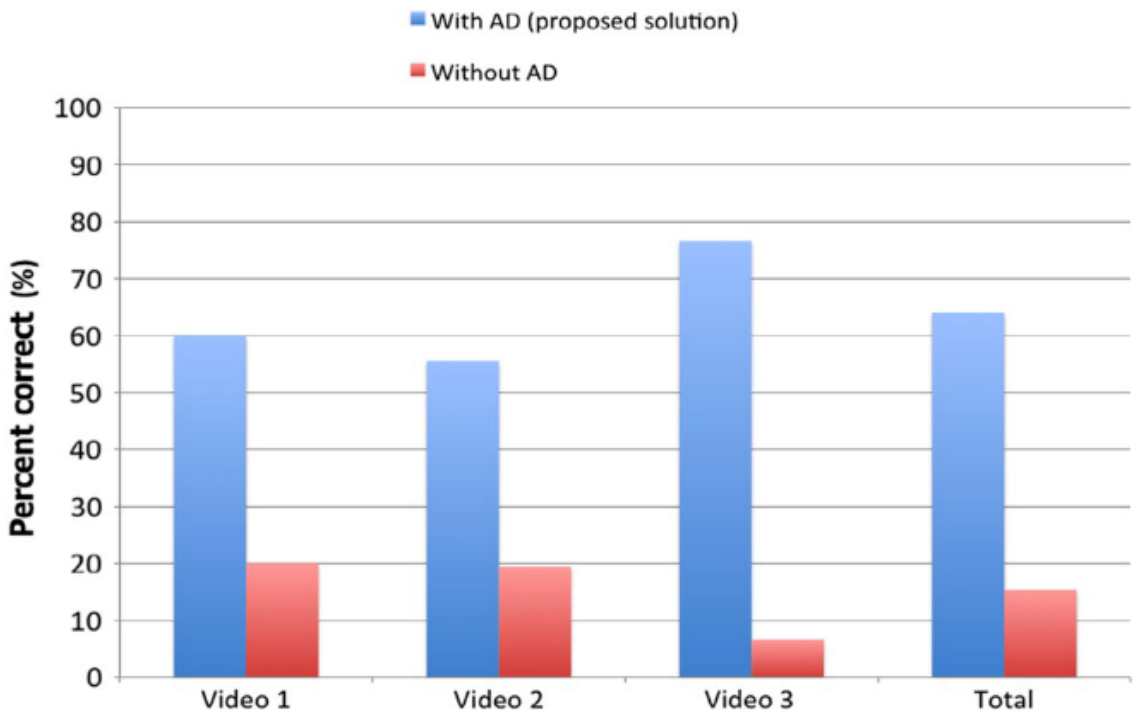

Abbildung 12: Korrekt beantwortete Fragen des Verständnis-Tests (Campos et al., 2020)

Im weiteren Schritt der Studie wurde der automatisch generierte Text von fünf Fachleuten analysiert und bewertet. Dabei stellte sich heraus, dass die Lückenerkennung sehr gut funktioniert hat, die Qualität der Texte jedoch zum Teil mangelhaft waren, da beispielsweise wichtige Elemente fehlten. Kritisiert wurde außerdem die Missachtung von offensichtlichen Soundeffekten und die an manchen Stellen fehlende Synchronität zwischen den Ereignissen im Film und der Audiodeskription. Das Fazit der Experten ist, dass der Einsatz des Systems als Referenz sinnvoll sein könnte, die Texte jedoch anschließend von Menschen modifiziert werden müssen (Campos et al., 2020).

Ein anderer, kreativer Ansatz, wurde 2011 von Viswanathan vorgeschlagen. Es handelt sich um einen Gürtel, der von den Sehbeeinträchtigten während des Films getragen wird und welcher ihnen durch Vibrationen Informationen zur Zuordnung der Personen auf dem Bildschirm vermittelt. Zur Umsetzung dieser Idee sind sechs Taktgeber im Halbkreis auf dem Gürtel positioniert und ein taktiler Rhythmus gibt die Position und die relative Entfernung der Figur vor der Kamera an. Visuell dargestellt ist dies in Abbildung 13.

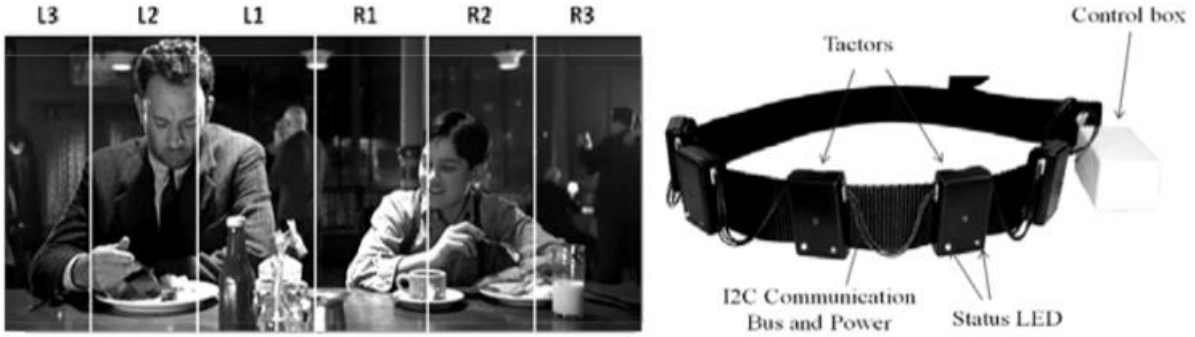

Abbildung 13: Zuordnung der Bildschirmregionen zu den Taktgebern auf dem Gürtel (Viswanathan et al., 2011)

Wenn beispielsweise nur der am weitesten links liegende Taktgeber vibriert, befindet sich die Person im Film ganz links im Bild. Für die Entfernung werden verschiedene Rhythmen angewandt. Wie in Abbildung 14 aufgezeigt, entspricht eine gleichmäßige Vibration einer geringen Entfernung der Person zur Kamera, wohingegen zwei kurze Vibrationen eine weite Entfernung angeben. Die lange Folge von kurzen Impulsen vermittelt eine mittlere Entfernung.

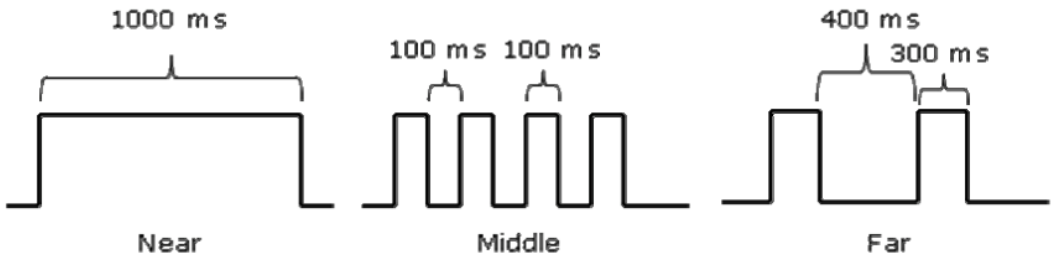

Abbildung 14: drei Rhythmen zur Angabe der Entfernung der Person zur Kamera (Viswanathan et al., 2011)

Getestet wurde diese Methode mit 14 Teilnehmenden anhand acht dreiminütiger Filmclips.

Die Resultate ergaben eine gute Einschätzung der Positionen der Personen, jedoch eine mangelhafte Einschätzung der Entfernung der Personen zur Kamera. Auch die Bewegungen der Personen waren schwierig nachzuvollziehen.

Trotzdem wurde dieser Ansatz von den Testenden als eine Bereicherung angesehen.

5.4 Personalisierung von Videoplayern

Ein barrierefreier Zugang zu Filmen und Videos ausschließlich im Fernsehen ist längst nicht mehr ausreichend, denn der Konsum von Multimedia-Inhalten weitet sich immer stärker auf das Internet aus. Filme werden heutzutage vermehrt auf Streaming-Plattformen oder in Mediatheken geschaut. Außerdem werden Videoinhalte zur persönlichen Weiterbildung, E-Learning, sowie für soziale Medien produziert. Um diese Inhalte für alle Menschen zugänglich zu machen werden Videoplayer benötigt, die einen barrierefreien Zugang ermöglichen (Moreno et al., 2017). Die WCAG-Richtlinien sehen dafür eine Bereitstellung von Audiodeskriptionen, Untertiteln, Transkriptionen und einer Gebärdensprache vor. Außerdem werden Erfolgskriterien beispielsweise zur Farbverwendung, der Kontrasteinstellung und der Textgröße aufgeführt *(Richtlinien für barrierefreie Webinhalte (WCAG) 2.1,* o. J.). Diese verschiedenen Einstellungsmöglichkeiten stellen eine große Herausforderung im Bereich der Gestaltung des UIs dar. Ein langes und komplexes Menü sollte vermieden werden (Brescia-Zapata, 2022).

Um die Barrierefreiheit eines Players zu testen, wird der BITV-Test durchgeführt und somit die Einhaltung der europäischen Norm (EN) 301 549 für barrierefreie Informations- und Kommunikationstechnik überprüft. Zum Bestehen des Tests wird eine Verfügbarkeit von dynamisch zuschaltbaren, textbasierten Untertiteln (closed captions) gefordert, die im Unterschied zu „open captions“ ein- und ausgeblendet werden können. Außerdem muss eine Audiodeskription eingebunden werden können, die entweder über ein Bedienelement aktiviert werden kann oder mittels einer separaten Version mit Audiodeskription im unmittelbaren Kontext des

Videos angeboten wird. Sowohl die Untertitelung als auch die Audiodeskription muss synchron zum Bild und Ton des Videos wiedergegeben werden können. Bei einer Übertragung, Konvertierung oder Aufnahme des Videos ist es erforderlich, dass die Untertitel bzw. die Audiodeskription erhalten bleiben und weiterhin synchron dargestellt werden. Des Weiteren sollten sich die Untertitel anpassen lassen. Es werden diesbezüglich Beispiele wie die Einstellungsmöglichkeit der Schriftgröße, der Schriftart oder der Hintergrundfarbe der Untertitel genannt, jedoch keine konkreten Vorgaben gemacht. Wenn die Sprache des Videos nicht der Hauptsprache des Webangebots entspricht, müssen zuschaltbaren bzw. programmatisch ermittelbare Untertitel in der Hauptsprache zusätzlich als akustische Ausgabe zugänglich gemacht werden. Beim Vorliegen von reinen Audiodateien oder stummen Videos bedarf es zudem einer gleichwertigen Medienalternative.

Zum Erfüllen der Anforderungen des Tests ist die Bedienbarkeit mit Tastatur, sowie eine barrierefreie Aktivierung der Zugänglichkeitsdienste durch die Nutzergruppe vorausgesetzt. Die Bedienelemente der barrierefreien Funktionen müssen sich dabei auf der gleichen Interaktionsebene, wie die Bedienelemente der Wiedergabenkontrolle (Play/Stopp, Lautstärke) befinden (*Prüfschritte BITV-Test / EN 301 549 (Web)* | *BIK BITV-Test Ergebnisse und Methodik* | *BIK BITV-Test*, o. J.).

Um den technischen Stand im Hinblick auf Barrierefreiheit und Personalisierungsmöglichkeiten von Videoplayern aufzuzeigen werden im Folgenden drei Player miteinander verglichen, die die Anforderungen der Norm weitestgehend erfüllen. Betrachtet werden der weltweit vielbenutzte YouTube-Player, der OZ-Player, der damit wirbt, das Konformitätslevel AA der WCAG 2.0 Richtlinien zu erfüllen und der Able-Player, der an der Hochschule der Medien im Rahmen der Master-Thesis von Remo Schneider entwickelt wurde.

Zu den Anforderungen der EN 301 549 werden in Tabelle 2 weitere Aspekte der WCAG-Richtlinien aufgegriffen und die Unterstützung dieser durch die Player untersucht. Das Verfügen bzw. das Fehlen der Funktionen ist in der Tabelle mit einem Haken oder einem Kreuz dargestellt.

Tabelle 2: Zugänglichkeit und Personalisierung, YouTube-Player, OZ-Player und Able-Player im Vergleich (Able Player Demos, o. J.; „Barrierefreiheit von YouTube“, 2018; OzPlayer, o. J.; Raemont, 2023; YouTube, o. J.)

	YOUTUBE-PLAYER	OZ-PLAYER	ABLE-PLAYER
Untertitel (CC)	✓	✓	✓
Verwenden von WebVTT-Dateien	✓	✓	✓
Einstellungsoptionen zu den Untertiteln	✓	✗	✓
Automatische Untertitelgenerierung	✓	✗	✗
Vorlesen von Untertiteln	(✓)	✗	✗
(De-) aktivierbare Audiodeskription	✓	✓	✓
Erweiterte Audiodeskription	✗	✗	✓
Synthetisierte Sprache für AD	✗	✗	✓
Einstellungsoptionen zur AD	✗	✗	✓
Transkription	✓	✓	✓
Gebärdensprache / multiple Videos	✗	✗	✓
Bedienbar mit Tastatur	✓	✓	✓
Ändern der Abspielgeschwindigkeit	✓	✗	✓
Einstellung der Bildqualität	✓	✗	✗
Ändern der Audiosprache	✓	✗	✗

Neben den Standardfunktionen wie Play/Stop, Lautstärke und Vollbildmodus, können alle drei Player mit der Tastatur bedient werden und bieten die Möglichkeit von Untertiteln. Die Bereitstellung einer Audiodeskription und einer Transkription ist ebenfalls bei allen drei Player möglich, jedoch ist die Einbindung einer Audiodeskription und Transkription bei einem YouTube-Video für den Laien etwas komplizierter, während die drei Dienste beim OZ-Player unmittelbar beim Hochladen des Videos durch Eingabe-Felder für Dateien eingebunden werden können.

Der Able-Player verfügt als einziger der drei Player über die Option einer erweiterten Audiodeskription, welche durch das Pausieren des Videos unterschiedlich ausführliche Beschreibungen ermöglicht. Da bei diesem Player die TTS-Technologie für die AD verwendet wird, existieren ebenfalls Auswahlmöglichkeiten hinsichtlich der Sprecherstimme, der Tonhöhe und der Geschwindigkeit der Sprache (*Able Player Demos*, o. J.; *OzPlayer*, o. J.; *YouTube*, o. J.).

Die Anpassung von Untertiteln in Form einer Farb-, Schriftgrößen- und Schriftart-Änderung ist sowohl beim YouTube-Player als auch beim Able-Player möglich. YouTube verfügt sogar über eine automatische Generierung und Übersetzung der Untertitel. Außerdem lassen sich die Untertitel durch das Hinzufügen der Erweiterung „Speak Subtitles for YouTube“ im Chrome-Browser mithilfe von synthetisierter Sprache vorlesen (Chrome Web-Store, 2022). Der OZ-Player verwendet zwar wie die anderen beiden Player ebenfalls WebVTT-Dateien für die Untertitelung, jedoch verfügt die graphische Benutzeroberfläche über keine dieser obengenannten Anpassungsmöglichkeiten. Lediglich das Ein- und Ausschalten der Untertitel ist möglich. Entgegen den Erwartungen ist das Ändern der Abspielgeschwindigkeit bei diesem Player jedoch nicht möglich, wohingegen der YouTube-Player, der Able-Player und viele weitere Player im Netz diese Funktion bereitstellen.

Die Option von multiplen Videos, welche die Verfügbarkeit eines Videos mit Gebärdensprache ermöglicht, wird von den hier aufgeführten Playern ausschließlich vom Able-Player bereitgestellt (*Able Player Demos*, o. J.). Bis vor Kurzem war außerdem bei keinem der Video-Player die Funktion des Umschaltens zwischen verschiedenen Audiosprachen eingebunden. Erst seit Anfang diesen Jahres ist die Erweiterung mehrerer Sprachen für das Video hinterlegen zu können auf der YouTube-Plattform möglich (Raemont, 2023).

Hinsichtlich der Zugänglichkeit für Menschen mit Beeinträchtigung sind der OZ-Player und der Able-Player eine sinnvolle Wahl, da sie sowohl Untertitel, eine Audiodeskription als auch eine Transkription zur Verfügung stellen können. Der YouTube-Player ermöglicht diese Funktionen ebenfalls, jedoch ist die Einbindung der Dienste, außerhalb der Untertitel, für laienhafte Videoproduzenten kompliziert. Darüber hinaus lässt sich der YouTube-Player mittels Screenreader schwieriger bedienen. Das Überspringen der Werbung ist beispielsweise für Blinde nicht möglich („Barrierefreiheit von YouTube“, 2018). Jedoch besitzt der YouTube-Player eine Menge weiterer Personalisierungsmöglichkeiten, die die anderen Player nicht bieten. Hierzu zählt die Möglichkeit eine Altersbeschränkung für die Videos hinzuzufügen, die Option die Bildqualität zu verändern, sowie die optionale Aktivierung des Autostarts einzustellen. Außerdem ist eine optionale Kommentarfunktion unterhalb der Videos geboten (*YouTube*, o. J.).

Obwohl inzwischen eine Menge Personalisierungen bei Videoplayern möglich sind, bedauert Brescia-Zapata (2022), dass bei der technischen Weiterentwicklung und der Produktion neuer multimedialer Inhalte das Thema der barrierefreien Zugänglichkeit nach wie vor ein nachträglicher Gedanke ist.

6 Praktischer Teil

Auf Grundlage der technologischen Recherche und durch das Erkennen von Problematiken der traditionellen Audiodeskription wurde in der vorliegenden Thesis ein neuer Ansatz zur personalisierten Audiodeskription entwickelt, dessen Anforderungen im Folgenden erläutert werden. Zudem wird die dazugehörige Feldstudie einschließlich Vorbereitung, Durchführung und Auswertung in diesem Kapitel beschrieben.

6.1 Anforderungen an den neuen Ansatz

In den vorherigen Kapiteln ist klar geworden, dass eine Audiodeskription vielen unterschiedlichen Wünschen und Anforderungen gerecht werden muss.

Dass eine Personalisierung im Bereich der Audiodeskription durch neue technologische Errungenschaften möglich ist, zeigen die bereits in dieser Arbeit aufgeführten Erweiterungen der klassischen Audiodeskription. Hierzu zählt die Anpassung der Sprechgeschwindigkeit als eine Lösung des Konflikts der unterschiedlichen auditiven Aufnahmevermögen von Menschen, sowie das optionale Pausieren des Videos, um eine ausführliche Beschreibung einfügen zu können (Minutella, 2022; Szarkowska, 2011). Bei Letzterem kann jedoch die Inhaltsmenge nicht gewählt werden.

Das in dieser Arbeit aufgeführte und getestete System zielt darauf ab, diese beiden Ansätze zu vereinen. Es bietet den Nutzenden unterhalb des Videoplayers eine Geschwindigkeit der Audiodeskription auszuwählen, welche gleichzeitig mit der Ausführlichkeit der Inhaltsbeschreibung zusammenhängt. Die Abhängigkeit gestaltet sich folgendermaßen:

Je schneller die gewählte Geschwindigkeit, desto ausführlicher erfolgt die Audiodeskription bzw. je langsamer die Geschwindigkeit, desto weniger Inhalt wird in der gleichen Zeit vermittelt. So kann das Stoppen des Videos für eine ausführlichere Beschreibung umgangen werden und je nach Bedürfnis eine langsamere Version gewählt werden. Die Differenziertheit der Geschwindigkeitsstufen beschränkt sich auf drei verschiedene Optionen, damit eine Übersichtlichkeit beibehalten werden kann. Die Stufen sind „langsam“, „normal“ und „schnell“ benannt. Die „normale“ Stufe entspricht dabei, sowohl hinsichtlich der Geschwindigkeit als auch bezüglich des Inhalts und der Wortwahl, der originalen menschlichen Audiodeskription.

Im Folgenden werden die Anforderungen an den Prototyp des Videoplayers und an das System mit entsprechendem Lösungsansatz aufgelistet und in *Abbildung 15* und *Abbildung 16* unterhalb der Liste veranschaulicht:

1. Es muss möglich sein, zwischen den drei Geschwindigkeiten und somit der Inhaltsmenge der Audiodeskription zu wechseln.
   Lösung: Ein Slider unterhalb des Videos (siehe *Abbildung 15*).
2. Eine eigenständige Bedienung durch die Testpersonen muss garantiert sein.
   Lösung: Eine barrierearme Webseite, auf der die Links zu den Videos hinterlegt sind (siehe *Abbildung 16*).
3. Zur Kostenreduktion ist eine günstige und einfache Herstellung der großen Menge an Audiomaterial für die drei verschiedenen Versionen der Audiodeskription sinnvoll.
   Lösung: Die Verwendung einer künstlichen Stimme.
4. Ein Vergleich zwischen der künstlichen Stimme und einer menschlichen Stimme muss herangezogen werden können.
   Lösung: Das Video zuerst mit menschlicher originaler Stimme zeigen und danach das System mit der künstlichen Stimme zum gleichen Filmausschnitt testen lassen (siehe *Abbildung 16)*.
5. Die Eignung für drei verschiedene Genres muss überprüft werden können.
   Lösung: Drei Videos zu den Genres Drama-, Action- und Dokumentarfilm mit je dreiminütigen Filmausschnitten auswählen (siehe *Abbildung 16)*.
6. Eine realistische und aussagekräftige Wahl von Probanden treffen.
   Lösung: sechs blinde oder stark sehbeeinträchtige Menschen, die potenzielle Nutzer des Systems sind.

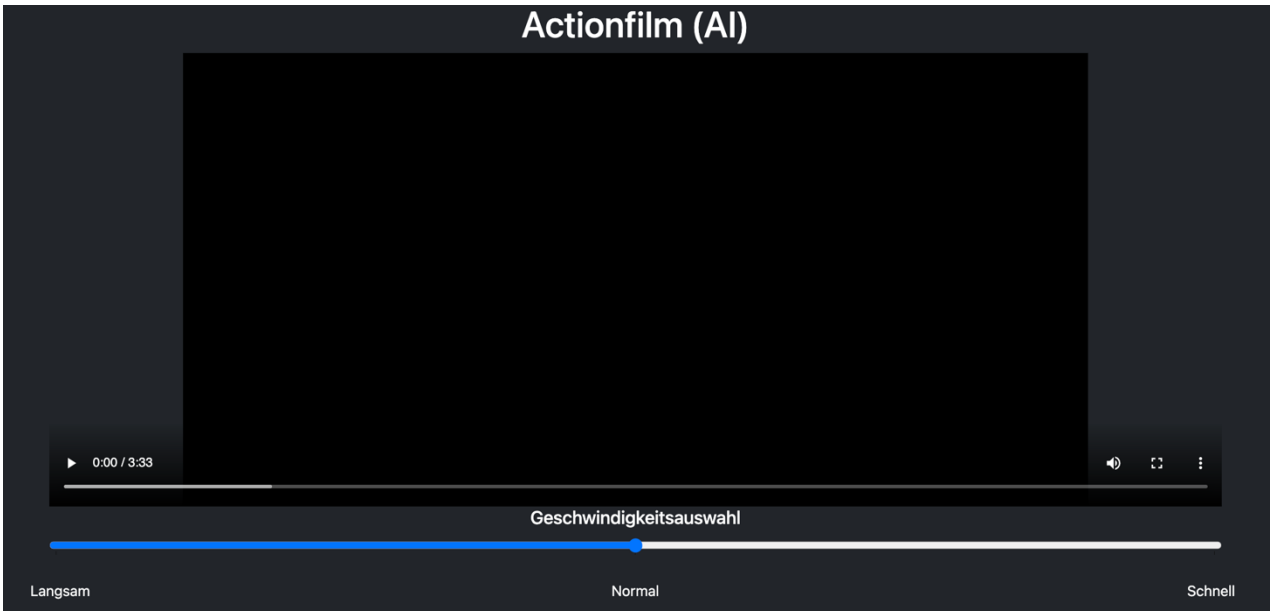

Abbildung 15: Benutzeroberfläche des Videoplayers mit wählbarer Geschwindigkeit (Übersicht — Personalisierte Audiodeskription, o. J.)

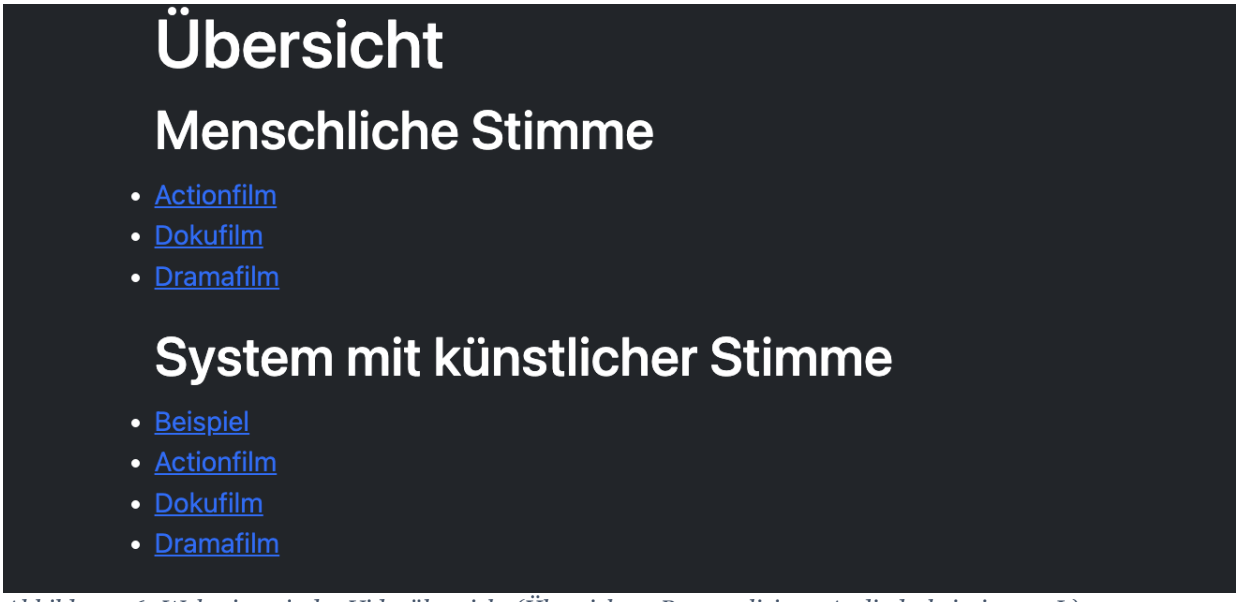

Abbildung 16: Webseite mit der Videoübersicht (Übersicht — Personalisierte Audiodeskription, o. J.)

6.2 Usability Test – Vorbereitung

Die Hauptaufgaben der Vorbereitung des Usability-Tests beziehen sich auf die Programmierung des Prototyps, die Wahl der Filmausschnitte und die entsprechende Produktion der AD-Texte, sowie die Rekrutierung der Testpersonen. Das Vorgehen und die Bearbeitung dieser Aufgaben werden in diesem Kapitel ausgeführt.

6.2.1 Programmierung des Prototyps

Für die Programmierung des Prototyps haben wir einen standardisierten HTML5-Player mit JavaScript erweitert und in eine Webseite eingebettet. Um die Funktion der Geschwindigkeitswahl zu ermöglichen, haben wir unterhalb der Standardfunktionen einen Slider hinzugefügt, durch den die Nutzenden zwischen den Optionen „langsam“, „normal“ und „schnell“ hin und her schalten können. Jeder Option ist ein eigenes Video mit einer eigenen Audiospur hinterlegt. Da für die zu testende Anwendung nur ein Wechsel der *Audiospur* notwendig ist, wird das gleiche Video verwendet und nur die Audiospur enthält entweder eine schnellere oder eine langsamere Version der Audiodeskription. Beim erstmaligen Öffnen der Seite haben wir die „normale“ Version voreingestellt. Sobald die Nutzenden umschalten, wird mittels JavaScript ein Videowechsel zur gewählten Option verursacht, wobei das Video am selben Punkt in der Zeitleiste weiterläuft.

Wir haben bei der Programmierung die Richtlinien des EN 301 549 und der WCAG für barrierefreies Webdesign weitestgehend eingehalten. Besonders haben wir darauf geachtet, eine

Zugänglichkeit für unsere Zielgruppe der Blinden und Sehbeeinträchtigten zu gewährleisten. Somit ist eine Bedienung für diese Menschen mithilfe eines gängigen Screenreaders, wie Voice-Over, JAWS oder NVDA, ohne Weiteres möglich.

Im folgenden Text wird exemplarisch der Codes der Actionfilm-Seite gezeigt.

```
<!DOCTYPE html>
<html lang="de">
<head>
    <meta charset="UTF-8">
    <meta name="viewport" content="width=device-width, initial-scale=1.0">
    <link rel="stylesheet" href="/main.css">
    <title>Actionfilm (AI) – Personalisierte Audiodeskription</title>
    <script>
      const versions = [
        {
          speed: 1,
          src: 'ai/action/1.mp4',
          type: 'video/mp4',
        },
        {
          speed: 2,
          src: 'ai/action/2.mp4',
          type: 'video/mp4',
        },
        {
          speed: 3,
          src: 'ai/action/3.mp4',
          type: 'video/mp4',
        },
      ];
    </script>

</head>
<body class="text-bg-dark text-center">
  <div class="container">

    <div class="row">
        <h1>Actionfilm (AI)</h1>
        <video id="video" class="ratio ratio-16x9" preload controls>
          <source type="video/mp4" src="/media/ai/action/2.mp4"/>
        </video>
        </div>
        <h5>Geschwindigkeitsauswahl</h5>
        <div class="row">
        <label for="speed">
          <span class="sr-only">Geschwindigkeitsauswahl</span>
          <input type="range" id="speed" name="speed" min="1" max="3" list="speed-values">

      </label>
        <datalist id="speed-values">
```

```
48.           <option value="1" label="Langsam"></option>
49.           <option value="2" label="Normal"></option>
50.           <option value="3" label="Schnell"></option>
51.       </datalist>
52.       </div>
53.       </div>
55.       <script src="/main.js"></script>
56. </body>
57. </html>
```

6.2.2 Videowahl

Bereits Kobayashi et al. haben ermittelt, dass künstliche Stimmen nicht bei jeder Filmsorte gleich gut funktionieren (Campos, 2020; Walczak, 2018).

Daher haben wir in dieser Studie drei möglichst unterschiedliche Genres getestet.

Bezüglich der Auswahl des Films innerhalb des Genres haben wir vor allem darauf geachtet, dass möglichst viele, für die Filmsorte charakteristischen Merkmale in dem gewählten Filmausschnitt auftreten. Für den Dramafilm-Ausschnitt bedeutet dies, dass Themen wie Konflikte im Privatleben der Hauptcharaktere behandelt werden, emotionale Situationen auftreten und ein erzählender Stil vorherrschend ist. Einen passenden Filmausschnitt haben wir in der Serie „In aller Freundschaft“ in der Folge „Zündstoff“ gefunden (*ARD Mediathek*, o. J.). Typische Merkmale eines Action-Films sind schnelle, möglicherweise parallele Handlungen, bei denen Spannung erzeugt und gegebenenfalls Gewalt gezeigt wird. Zur Aufzeigung dieser Merkmale haben wir einen Ausschnitt der Serie „Mirage“ aus der Folge „Geister der Vergangenheit“ gewählt (*ZDF Mediathek*, o. J.). Als Dokumentation haben wir die Naturdokumentation „Wildes Kalifornien – Ströme des Lebens“ ausgesucht, bei der wir jedoch einen Ausschnitt mit Erklärungen zu möglichst allgemeingültigen wissenschaftlichen Phänomenen gewählt haben (*ARD Mediathek*, o. J.). Dadurch lassen sich die Ergebnisse unter Umständen auch auf andere Dokumentationsarten übertragen.

Bei der Auswahl der Filmsequenzen haben wir nicht zwischen Serien und Filmen differenziert, da dies für dreiminütige Ausschnitte keine Relevanz trägt.

6.2.3 Erstellung der AD-Skripte und Synthetisierung in Sprache

Die Beschreibung der erweiterten Versionen haben wir anhand vorheriger, ausführlicher Recherche selbst konzipiert. Die originale menschliche Audiodeskription nahmen wir hierfür als Referenz.

Für die Erstellung der schnellen Version haben wir beispielsweise Informationen zur Kameraführung, zur Farbgestaltung, zur Mimik und Gestik der Personen und Umgebungsbeschreibungen hinzugefügt.

Bei der langsamen Version hingegen haben wir Informationen, die nicht zwingend zur Verständnisvermittlung beigetragen haben, weggelassen.

Die geschriebenen Skripte haben wir in einer Excel-Tabelle erfasst und schließlich über die Cloud-Computing-Plattform „Microsoft-Azure“ von Text in Sprache synthetisieren lassen. Für diese Synthetisierung stehen dort sieben männliche und acht weibliche neuronale Stimmen für die deutsche Sprache zur Verfügung. Den Text haben wir auf einer graphischen Benutzeroberfläche eingefügt und anschließend hinsichtlich einiger Merkmale wie Sprachgeschwindigkeit, Intonation, Aussprache und Pausensetzung angepasst. Das UI (User Interface) ist in Abbildung 17 zu sehen.

Das Exportieren der synthetisierten Sprache erfolgte mittels einer Wavedatei, welche wir in der DAW Cubase auf die Videos zugeschnitten haben.

Der letzte Schritt der Vorbereitungsphase in der Inhaltserstellung war die Einbindung der verschiedenen Videos in den auf der Webseite liegenden Player.

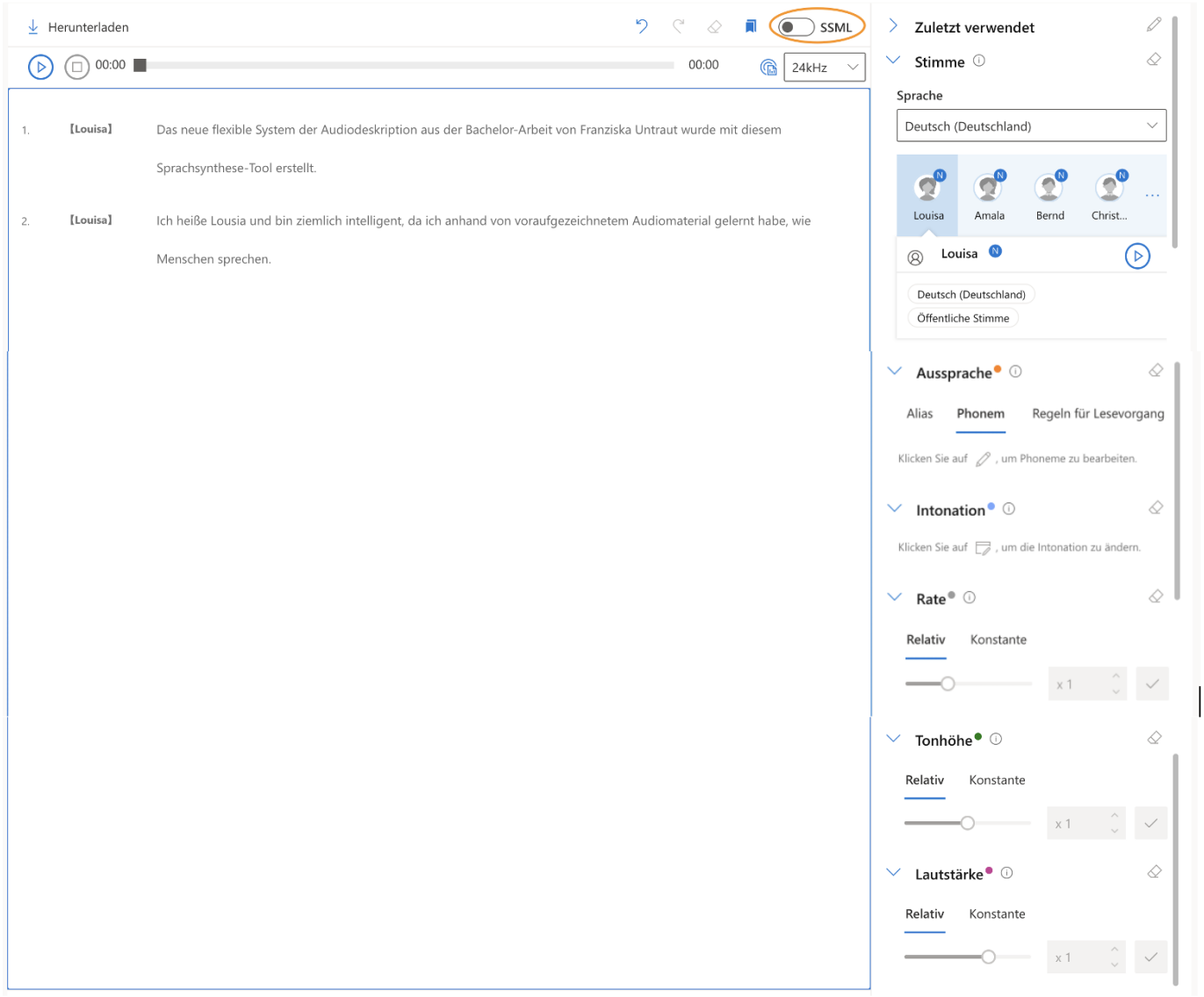

Abbildung 17: Microsoft Azure als Software-Tool zur Erstellung synthetisierter Sprache (Speech Studio - Microsoft Azure, o. J.)

6.2.4 Testpersonen

Für eine Teilnahme an der Studie mussten die Testpersonen folgende Kriterien erfüllen: Es musste eine starke Sehbeeinträchtigung oder Blindheit vorliegen, die untere Altersgrenze betrug 18 Jahre und ein PC oder Laptop musste eigenständig mittels Screenreader bedient werden können.

Alle Personen mussten außerdem einwilligungsfähig sein und muttersprachlich deutsch sprechen.

Wir haben sechs Personen rekrutiert, welche wir über die Vermittlung von Professoren der Hochschule der Medien, sowie über das Kontaktieren eines Hörfilm- und Blindenvereins erreicht haben.

Die Probanden waren zwischen 28 und 60 Jahre alt, vier davon männlich und zwei weiblich. Es waren nicht alle Probanden geburtsblind, sondern die Hälfte hat erst im Laufe ihres Lebens, überwiegend im Jugendalter, das Augenlicht verloren. Der Blindheitsgrad betrug jedoch bei allen Personen volle oder nahezu volle 100%. Außerdem hatten alle Teilnehmenden bereits Erfahrung mit Audiodeskription. Ein zufälliger, aber doch zu berücksichtigender Aspekt ist, dass zwei von den sechs Probanden beruflich in der Herstellung von Audiodeskriptionsskripten tätig sind.

6.3 Usability Test – Durchführung

Im anschließenden Teil werden die in der Studie angewendeten Methoden und der Verlauf der praktischen Testung des Prototyps beschrieben.

6.3.1 Methoden des Usability Tests

Bei der Studie handelt es sich um eine qualitative Feldstudie, in der wir eine geringe Anzahl an Probanden ausführlich befragt haben.

Wir haben sowohl offene Fragen als auch Fragen, die mithilfe einer Likert-Skala beantwortet werden konnten, gestellt. Das Niveau dieser Skalen bestand einheitlich aus fünf Merkmalsausprägungen. Ebenfalls haben wir einzelne Entscheidungsfragen aufgeführt, welche jedoch mittels weiterer Fragen ergänzt oder durch Begründungen ausgeführt wurden.

Den Usability-Test und das darauffolgende Interview haben wir mit jeder Person einzeln durchgeführt. Die Daten der Personen wurden anonymisiert und das Gespräch zu Nachweiszwecken mit vorherigem Einverständnis aufgezeichnet.

Als Methode des Testverlaufs haben wir das Prinzip des „Within-Subject-Designs“ angewendet. Daraus resultiert, dass jeder Teilnehmende beide Systeme testet und danach miteinander vergleicht.

Zuerst haben wir den Testpersonen die originale menschliche Audiodeskription abgespielt, dann das System mit der variablen künstlichen Stimme.

Da der Videoplayer von den Testpersonen eigenständig bedient wurde, konnten diese beim neuen System selbst entscheiden, ob sie die drei Deskriptionen nacheinander anschauen, oder während des Abspielens hin und her wechseln. Die Reihenfolge der Videos hingegen, war von uns festgelegt, um mittels Counterbalancing dem Einfluss eines möglichen Lerneffekts, entgegenzuwirken.

Bei drei Genres mit sechs Testdurchläufen belegte daher jedes Genre genau zweimal jede Stelle der Abfolge.

Ersichtlich wird dies in folgender Tabelle:

Tabelle 3: Counterbalancing: Reihenfolge der Videos für die jeweiligen Testpersonen (eigene Darstellung)

	PERSON 1	**PERSON 2**	**PERSON 3**	**PERSON 4**	**PERSON 5**	**PERSON 6**
1. VIDEO	Drama	Drama	Doku	Doku	Action	Action
2. VIDEO	Action	Action	Drama	Drama	Doku	Doku
3. VIDEO	Doku	Doku	Action	Action	Drama	Drama

Um das Umschalten zwischen den Optionen zu erlernen, bediente jeder der Testpersonen am Anfang des Tests ein Beispielvideo.

Da der Test online erfolgte, gaben die Probanden zu Zwecken der Mitverfolgung ihren Bildschirm und das Audio frei. So konnten wir die „Think-Aloud-Methode“ in der Forschung anwenden. Der Videoausschnitt der menschlichen Stimme durfte lediglich einmal pro Genre angeschaut werden, das System mit der künstlichen Stimme maximal zehn Minuten. Somit war es den Testenden möglich, jede Geschwindigkeitsstufe einmal vollständig anzuhören.

Die Dauer des Interviews variierte zwischen zwanzig Minuten und einer Stunde, sodass die Gesamtdauer des Tests pro teilnehmende Person zwischen ein und zwei Stunden betrug.

6.3.2 Durchführung

Zu Beginn der Studie haben wir das Einverständnis zur Aufzeichnung und zu weiteren auf dem Informationsblatt aufgeführten Daten eingeholt. Danach erfolgte eine Erklärung zur neuentwickelten Methode mit künstlicher Stimme, sowie eine weitere Informationsgabe zum Ablauf des Tests. Die Links zu den anzuschauenden Videos wurden auf einer Webseite bereitgestellt, auf welche die Probanden Zugriff hatten (*Übersicht — Personalisierte Audiodeskription*, o. J.). Nachdem sie sich einen Überblick über diese Videoübersicht verschafft hatten, haben sie das neuartige System an einem Beispielvideo getestet. Das hat in der Regel gut funktioniert, doch wurde manchen Teilnehmenden erst währenddessen klar, wie das System grundsätzlich funktioniert. Daher war das Beispielvideo von großem Nutzen. Gewählt haben wir hierfür einen Ausschnitt des Märchens „Schneewittchen“, welcher viel Audiodeskription und wenig Dialog enthielt.

Nach dieser Einführungsphase startete der relevante Teil des Tests mit dem Anschauen des Videos des ersten Genres. Zuerst haben die Teilnehmenden die originale menschliche Audiodeskription, dann die Version mit der künstlichen Stimme und dem wählbaren System,

konsumiert. Einige Probanden haben bereits nach jedem Genre kommentiert, andere haben mit dem Feedback bis zum Interview gewartet.

Nicht selten waren kleine Geräusche und spontane Reaktionen während des Anschauens interessant und aufschlussreich.

Die Bedienung des Players war für die meisten blinden Menschen nicht ganz einfach, jedoch haben es alle, bis auf eine Person, selbstständig hinbekommen. Der einen Person, die aufgrund von Internetproblemen Schwierigkeiten hatte, haben wir die Videos mithilfe ihrer Anweisung vorgespielt.

Da es von uns keine Vorgabe gab, in welcher Reihenfolge und auf welche Art und Weise das System mit den wählbaren Geschwindigkeiten erkundet werden sollte, war die Vorgehensweise der Teilnehmenden interessant zu beobachten. Die meisten wählten zuerst die schnelle Geschwindigkeit, da sie neugierig waren, welche Informationen dazukommen. Dann haben sie auf die langsame Version geschaltet, wobei Einige diese Version nicht vollständig durchhörten. Die mittlere Stufe hat keiner komplett konsumiert, da der Text und die Geschwindigkeit dieselbe wie bei der zuvor gehörten menschlichen Beschreibung war und daher der Anreiz fehlte. Überraschenderweise hat nur einer der Teilnehmenden die Zeitleiste zum Vor- und Zurückspringen genutzt, um einen direkten Vergleich zwischen zwei Versionen einer beschriebenen Lücke zu erkunden. Die Vermutung liegt nahe, dass die Bedienung dieser Funktion etwas zu herausfordernd für die erste Nutzung des Systems war. Jedoch hatten wir nicht das Gefühl, dass dies notwendig gewesen wäre, denn alle Testenden konnten hinterher sehr genau sagen, was sich an den Informationen geändert hatte.

Im Anschluss folgte ein Interview, bei dem sowohl unsere vorbereiteten Fragen beantwortet wurden als auch umfassende Ausschweifungen zum Thema Audiodeskriptionen Platz gefunden haben.

6.4 Usability Test - Ergebnisse und Auswertung

Um das vorgeschlagene System besser bewerten zu können, werden die im Interview gestellten Fragen in fünf Kategorien strukturiert.

- **Neuronale Stimme**
- **Geschwindigkeitsstufen**
- **Genres**
- **Inhalt der AD-Texte**
- **Allgemeine Fragen zum System**

Im Folgenden werden die Ergebnisse pro Kategorie aufgeführt und anschließend bewertet.

Neuronale Stimme:

1. *Wie gut fanden Sie die Qualität der künstlichen Stimme? An welchen Aspekten muss noch gearbeitet werden?*
2. *Fanden Sie es schwierig die künstliche Stimme zu verstehen und den Inhalt aufzunehmen?*
3. *Wie stark haben die Eigenschaften der künstlichen Stimme die Immersion abgeschwächt?*
4. *Wie gut war die Emotionsvermittlung bei den Versionen mit der künstlichen Stimme?*

Alle vier Fragen wurden anhand einer Likert-Skala mit fünf Merkmalsausprägungen beantwortet und ihre Ergebnisse in Abbildung 18 dargestellt. Aus ihrem Mittelwert wurde die Qualität der neuronalen Stimme abgeleitet.

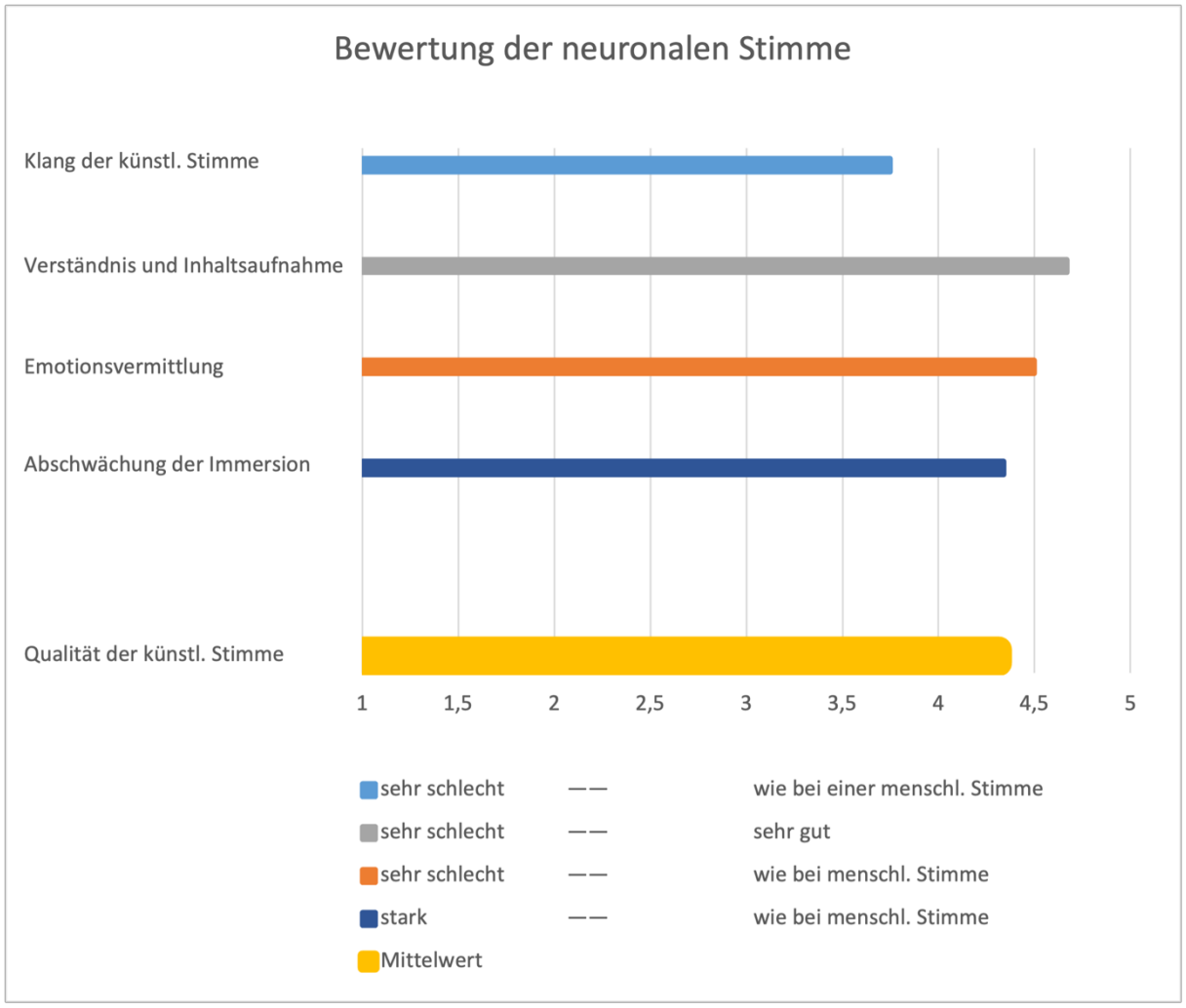

Abbildung 18: Ergebnisse zu den Fragen bezüglich der neuronalen Stimme (eigene Darstellung)

Der Wert „1“ entspricht der schlechtesten Bewertung, während eine „5“ die bestmögliche Bewertung beschreibt und mit der Qualität einer menschlichen Stimme gleichzusetzen ist.

Der Klang der Stimme wurde im Mittel mit 3,75 Punkten bewertet, wobei eine hohe Varianz mit einer Wertespanne zwischen 2 und 5 Bewertungspunkten, existiert. Während einige Testpersonen eine störende Eintönigkeit der künstlichen Stimme anmerkten, war ein anderer Proband der Meinung, keinen nennenswerten Unterschied zur menschlichen Stimme zu bemerken. Er fügte außerdem hinzu, dass die künstliche Stimme die positive Eigenschaft habe, trotz einer enorm hohen Sprechgeschwindigkeit eine sehr präzise Artikulation zu erzeugen, wie es kein Mensch bei dieser Geschwindigkeit schaffen würde.

Beim Verständnis und der Inhaltsaufnahme müssen durch die künstliche Stimme im Vergleich zur menschlichen Stimme kaum Abstriche gemacht werden. Die Testpersonen bewerteten diesen Bereich mit der Höchstpunktzahl von 4,7 Punkten.

Die Emotionsvermittlung beim Filmausschnitt mit der künstlichen Stimme wurde mit 4,5 Punkten bewertet und einige Testpersonen betonten, dass die Emotionen eines Films nicht durch die Audiodeskription, sondern durch die Dialoge, die Musik und die Soundeffekte vermittelt werden. Dieses Argument lässt auch nachvollziehen, warum die Abschwächung der Immersion, also des Eintauchens in die virtuelle Umgebung, durch die künstliche Stimme nur gering ist. Die Bewertung beträgt hier 4,3 Punkte, wobei die Punktzahl „1“ eine sehr starke Abschwächung und die Punktzahl „5“ keine Abschwächung bedeutet.

Aus der Errechnung des Mittelwertes dieser vier Ergebnisse, resultiert die Bewertung der Qualität der neuronalen Stimme von 4,3 Punkten. Daraus lässt sich eine hohe Zufriedenheit und Akzeptanz dieser künstlichen Stimme ableiten.

5. *Im Test wurde einheitlich eine weibliche Sprecherstimme verwendet. Denken Sie eine männliche Stimme mit gleicher Qualität hätte Ihre Empfindungen geändert?*

Auch wenn zwei der Teilnehmenden der Meinung sind, dass es einen deutlichen Unterschied zwischen einer männlichen und einer weiblichen Sprecherstimme gibt, sind alle der Meinung, dass aufgrund der einheitlichen Verwendung der weiblichen Stimme die Einschätzung des Systems durch diesen Aspekt nicht beeinflusst oder verfälscht wurde.

Geschwindigkeitsstufen:

6. *Welche Geschwindigkeitsstufe würde von Ihnen am häufigsten genutzt werden?*

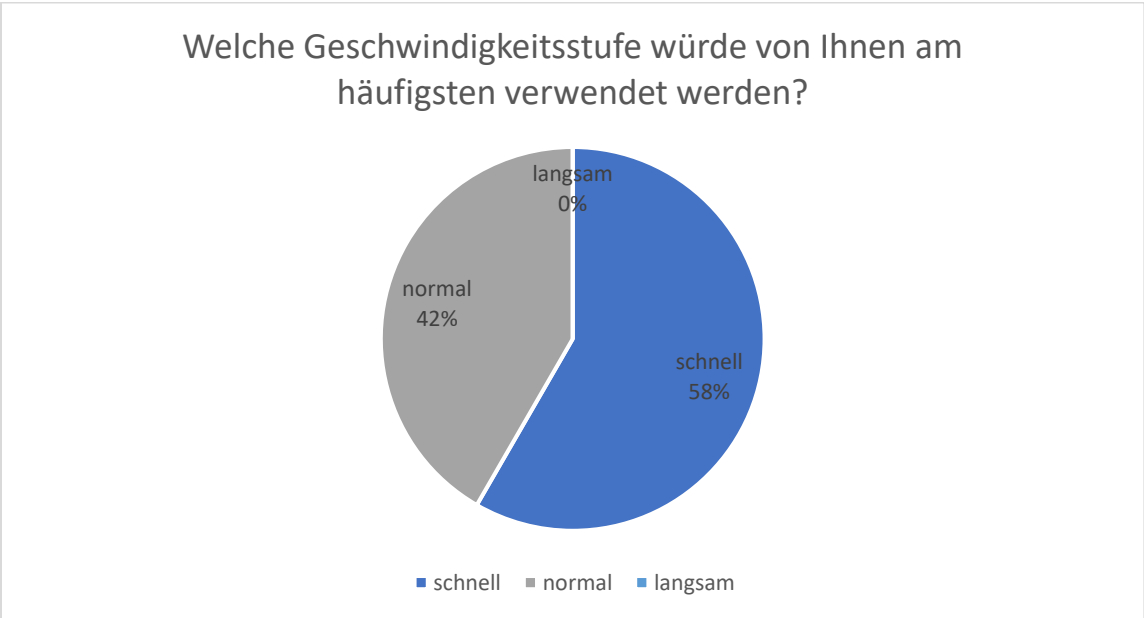

Abbildung 19: statistische Verteilung der Bevorzugung der Geschwindigkeitsstufen (eigene Darstellung)

Eine Menge der Testpersonen fokussierten sich während der Durchführung des Tests stark auf die schnelle Geschwindigkeitsauswahl. Daher leuchtet es ein, dass diese mit einem Wahlergebnis von 58% die beliebteste Geschwindigkeitsstufe verkörpert. Das bestätigt, dass definitiv ein Interesse an mehr Informationen besteht und dass die normale Geschwindigkeitsstufe, welche der menschlichen Audiodeskription entspricht, für manche Menschen möglicherweise zu wenig Inhalt bietet.

Die verbleibenden 42% Prozent verteilen sich auf die normale Geschwindigkeitsstufe. Das bedeutet, dass keiner die langsamste Stufe als bevorzugte Geschwindigkeitsstufe gewählt hat. Demnach sollte bei der Einführung eines solchen Systems überlegt werden, ob das Einbeziehen dieser Stufe einen Sinn ergibt.

Die Untersuchung einer Korrelation im Zusammenhang mit der Bevorzugung der Geschwindigkeitswahl wurde in Bezug auf das Alter und den Zeitpunkt der Erblindung vorgenommen. In Abbildung 20 ist das Alter der Testpersonen mit blauen Punkten und der Zeitpunkt der Erblindung mit orangenen Punkten dargestellt.

Die y-Achse gibt dabei die Werte beider Untersuchungsgegenstände an.

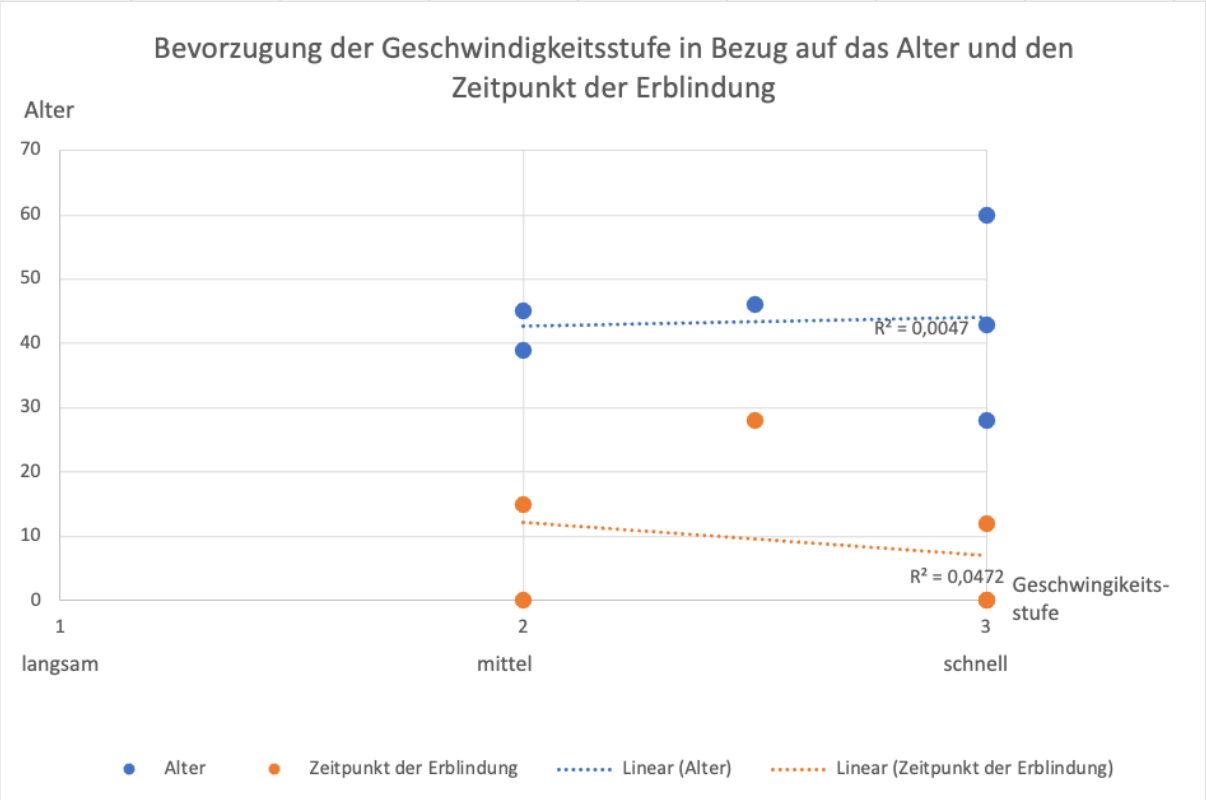

Abbildung 20: Untersuchung einer Korrelation zwischen der Geschwindigkeitswahl und des Alters bzw. des Zeitpunkts der Erblindung (eigene Darstellung)

In beiden Fällen lässt sich keine Korrelation feststellen, da das Bestimmtheitsmaß zu gering ist. Aufgrund der geringen Anzahl an Probanden ist das Ergebnis jedoch nicht signifikant, da bei der Änderung eines Punktes um eine Geschwindigkeitsstufe bereits eine Korrelation festgestellt werden kann.

7. *Wie haben Sie die Geschwindigkeit der langsamsten Stufe empfunden?*

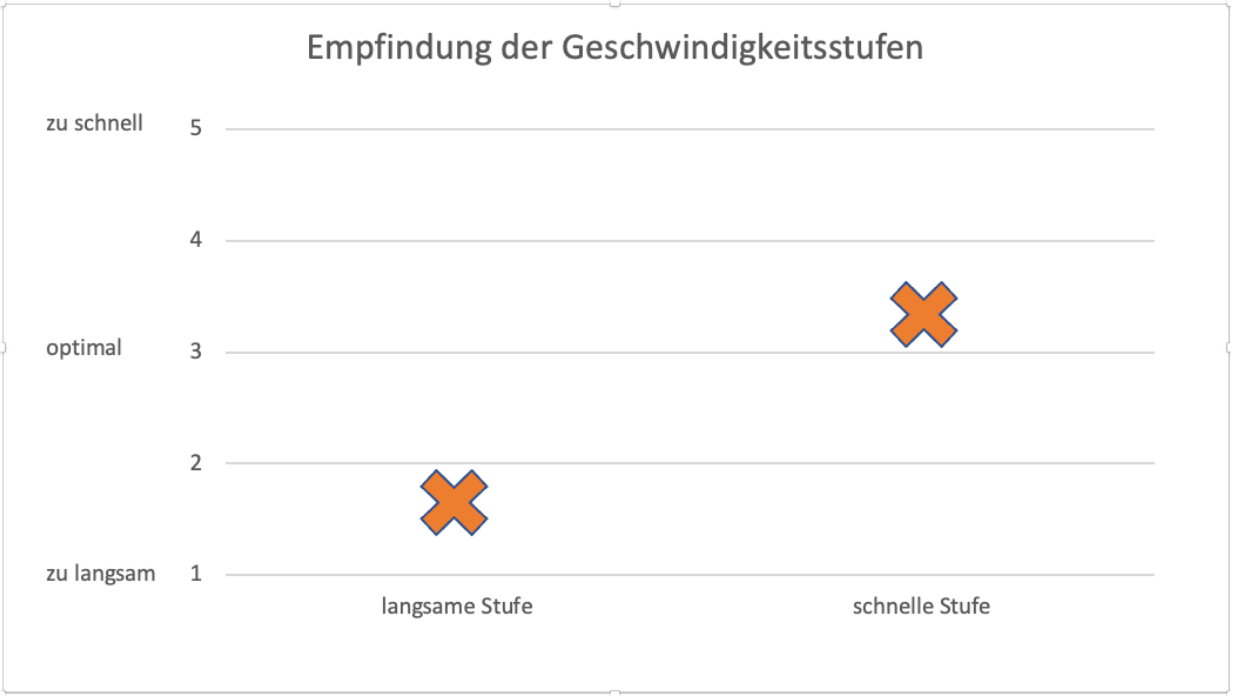

Abbildung 21: Empfindung der Geschwindigkeitsstufen langsam und schnell (eigene Darstellung)

Die persönliche Empfindung der schnellen und langsamen Stufe wurde mithilfe einer Likert-Skala bewertet, wobei „1“ einer zu langsamen Empfindung, „3“ einer optimalen Empfindung und „5“ einer zu schnellen Empfindung der Geschwindigkeit entspricht.

Wie in Abbildung 21 abgebildet, wurde die schnellste Stufe im Mittel mit einer Punktzahl von 3,3 Punkten bewertet und von einigen Testpersonen als optimal wahrgenommen. Die langsame Stufe hingegen war mit einer Punktzahl von gerade einmal 1,7 Punkten weit unter dem auditiven Aufnahmevermögen der meisten getesteten Personen. Einige kommentierten, dass sie die erste Stufe niemals verwenden würden, wohingegen eine Teilnehmerin von einigen ihrer Bekannten berichtete, die sicherlich gerne die langsame Stufe verwenden würden. Daher wäre es sinnvoll, das System mit einer größeren Studienpopulation, mit Menschen aus unterschiedlichen Bildungsschichten und Lebenssituationen zu testen. Meine Testpersonen besaßen alle weitgehendst einen hohen Bildungsgrad und dadurch möglicherweise ein hohes auditives Aufnahmevermögen.

8. *Wäre es hilfreich weitere Unterteilungen der Geschwindigkeit zu haben? Wenn ja, welche und wie viele?*

Trotz der Abneigung vieler Testpersonen gegen die langsame Stufe würden fünf von sechs Personen nichts an der Unterteilung ändern. Argumentiert wurde, dass die langsame Stufe sicherlich hilfreich für frisch erblindete Personen sein könnte, die sich erst an Audiodeskriptionen gewöhnen müssten und dass man selbst nicht der Maßstab für alle sei. Eine Person würde die langsame Stufe weglassen und stattdessen eine noch schnellere Version hinzufügen, falls ein Gewöhnungseffekt eintreten sollte. Drei verschiedene Inhaltsmengen sind demnach aber für alle Testpersonen ausreichend.

Ein weiterer Vorschlag eines Probanden ist es, die Geschwindigkeit der Sprache in Prozent anzugeben und ab einer bestimmten Prozentzahl die Inhaltsmenge zu ändern.

9. *Wie sinnvoll finden Sie die Funktion während des Abspielens die Geschwindigkeit wechseln zu können?*

Fünf der sechs Testpersonen finden die Funktion zwischen den Geschwindigkeiten wechseln zu können sinnvoll und würden sie selbst nutzen. Eine Begründung für die Nützlichkeit der Wechselfunktion ist zum einen, dass die Konzentration nach einiger Zeit nachlässt und man durch diese Funktion während des Films eine Geschwindigkeitsstufe langsamer schalten kann. Zum anderen wurde argumentiert, dass man Szenen für das Verständnis mit der ausführlichen Variante erneut angucken könnte, falls eine der langsameren Varianten keine zufriedenstellende Vermittlung liefert. Außerdem kann die Ausführlichkeitsstufe dem Interesse der einzelnen Szenen im Film angepasst werden. Ein Beispiel hierfür sind Kampfszenen, welches in den genrespezifischen Fragen näher erläutert wird.

Der Proband, der gegen das Wechseln der Geschwindigkeitsstufen während des Abspielens stimmte, argumentierte, dass eine Audiodeskription inhaltlich aufeinander aufbaut und ein in sich geschlossener Text ist. Der Wechsel von verschiedenen Versionen würde seiner Meinung nach zu Verwirrung führen, da in der schnellen Variante möglicherweise Dinge vom Anfang des Films im späteren Verlauf erneut aufgegriffen werden, die in der langsamen Version jedoch anfangs nie erwähnt wurden.

Mit Sicherheit stellt dies eine Herausforderungen für die Verfassenden der Texte dar und müsste anhand eines ganzen Films getestet werden.

10. *Finden Sie es sinnvoll, dass man die Geschwindigkeit an die Informationsmenge koppelt?*

Alle Teilnehmenden empfanden dies als eine logische Schlussfolgerung für das Erreichen einer erhöhten Vermittlung des Inhalts.

Genres:

11. *Erstellen Sie eine Reihenfolge der Genres, die sich Ihrer Meinung nach am besten für dieses System eignen.*

Um Platzierungen für die Eignung der Genres für das System zu erstellen, wurden die persönlich gewählten Reihenfolgen der Teilnehmenden folgendermaßen verrechnet.

Die Wahl des ersten Platzes erhielt den Punkte-Wert „3“, der zweite Platz den Wert „2 “ und der dritte Platz die Punktzahl „1“. Diese Wertungen wurden addiert, wobei anschließend das Genre mit der höchsten Summe den ersten Platz belegte. Mit absteigender Punktzahl erfolgte die Zuordnung der weiteren Platzierungen.

Anders als möglicherweise erwartet gibt es kein Genre, das klar favorisiert wird.

Der Actionfilm liegt mit einer Punktzahl von 14 Punkten auf Platz 1, der Dramafilm mit 12 Punkten auf Platz 2 und die Naturdokumentation mit 11 Punkten auf Platz 3.

In Abbildung 22 werden diese Platzierungen dargestellt, wohingegen in Abbildung 23 die Zusammensetzung der einzelnen Plätze aufgezeigt wird.

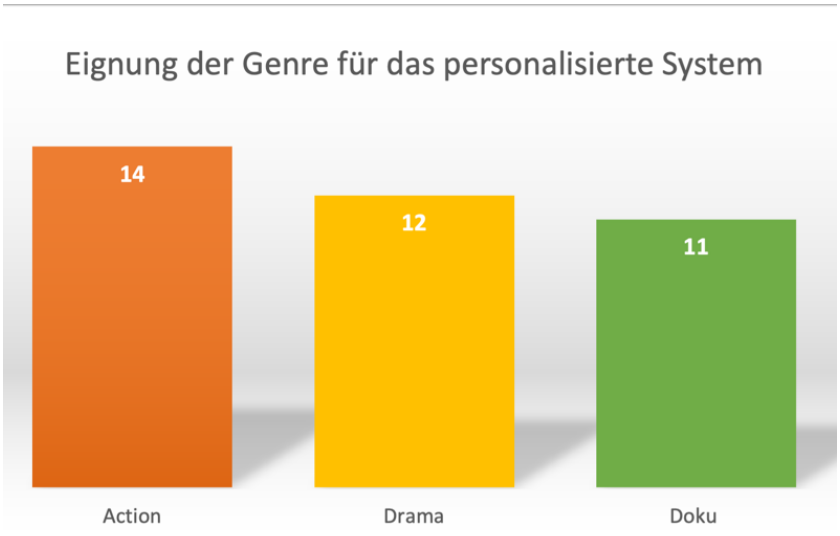

Abbildung 22: Platzierung der besten Eignung der Genres für das personalisierte System (eigene Darstellung)

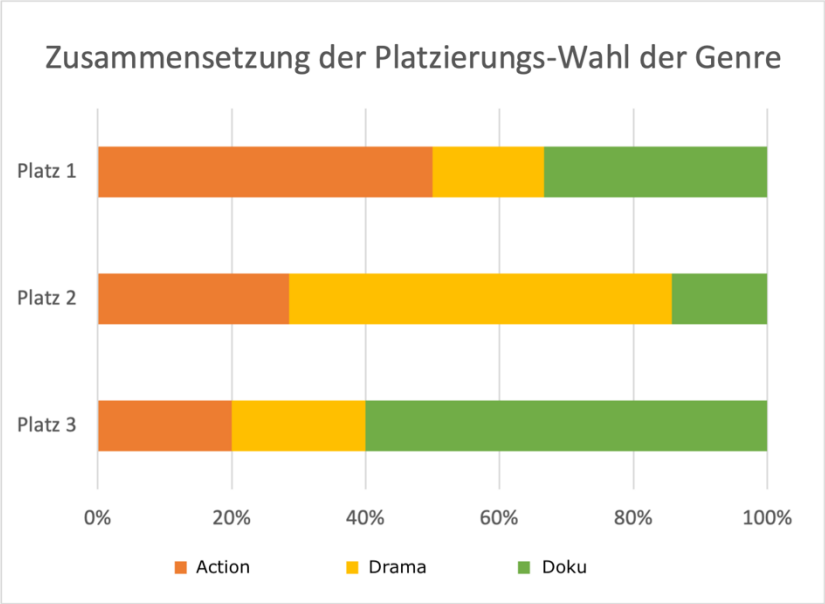

Abbildung 23: Zusammensetzung der Platzierungen pro Genre (eigene Darstellung)

Die zwei Hauptgründe, warum einige Testpersonen den „Actionfilm“ auf Platz 1 gesetzt haben, sind erst einmal, dass häufig schnelle Handlungen parallel stattfinden, welche zusätzlich oft sprunghaft geschnitten sind. Daher kann eine schnelle Beschreibung hilfreich sein. Zweitens gibt es Menschen, die nicht gerne Gewaltszenen sehen und deshalb die Wechselmöglichkeit der Ausführlichkeit der Beschreibung bei solchen Szenen gerne in Anspruch nehmen.

Bei der Zusammensetzung der einzelnen Plätze ist interessant zu beobachten, dass das Genre „Dokumentation“ besonders häufig auf Platz 1 oder Platz 3 gesetzt wurde, wohingegen der zweite Platz besonders häufig vom Genre „Drama“ belegt wird. Daraus lässt sich ableiten, dass die Sinnhaftigkeit des Systems bei einer Dokumentation stärker umstritten ist und die Meinungen polarisieren, während sich beim Drama die meisten Probanden das System vorstellen können, jedoch den Mehrwert dabei nur begrenzt einschätzen.

Kommentare der Testpersonen bestätigen diese Schlussfolgerungen. Ein Proband findet, dass die Erzählerstimme einer Dokumentation bereits die Hauptinformationen vermittelt und eine Personalisierung der Audiodeskription dadurch an Nützlichkeit und Notwendigkeit verliert. Demgegenüber steht die Meinung einer anderen Testperson, die empfindet, dass Dokumentationen von Bildern leben und je ausführlicher die Beschreibung ist, desto stärker wird diese visuelle Ebene transportiert. Dieser ergänzt, dass die Wechselfunktion bei dem Genre der „Dokumentation“ am sinnvollsten sei, da das Konsumieren einer Dokumentation häufig stark unterschiedliche Gründe hat. Manchmal werden Dokumentationen angeschaut, um zu entspannen und ein anderes Mal, um etwas zu lernen. Beim Genre „Drama“ wird angemerkt, dass häufig wenig Platz zwischen den Dialogen zu Verfügung steht und die Möglichkeit auf eine schnelle

Version wechseln zu können, nützlich sein könnte. Andererseits finden häufig alltägliche und einfache Handlungen statt, bei denen einige visuelle Elemente nicht relevant für das Grundverständnis sind und es nicht gravierend ist, wenn diese in der Audiodeskription weggelassen werden.

Zusammenfassend kann festgestellt werden, dass die Meinungen bezüglich der Eignung der Genres sehr subjektiv und unterschiedlich sind. Dadurch wäre eine Bereitstellung des Systems zu jedem der aufgeführten Genres sinnvoll, sodass zumindest die Option eines Wechsels der Beschreibung besteht.

12. Hätten Sie gerne ein anderes Genre getestet, wenn ja welches?

Instinktiv wurde diese Frage häufig zuerst mit „Nein“ beantwortet. Nach Überlegungen wurde vorgeschlagen die Testung dieses Systems an Sport- und Musikfilmen vorzunehmen, sowie bei der Live-Übertragung von Sportevents auszuprobieren. Der Reiz bei einem Musikfilm oder Musical besteht durch die Eingliederung der künstlichen Stimme in das Kunstwerk. Bei einem Sportfilm und Sportevents sind es die vielen Aktionen, die gleichzeitig geschehen und möglicherweise eine sehr schnelle Beschreibung benötigen.

Inhalt der AD-Texte:

13. AD dient vor allem der Verständnisvermittlung der Handlung. Wie wichtig ist es Ihnen, dass ebenfalls Emotionen transportiert und beschrieben werden?

Auf einer Likert-Skala mit fünf Wahlmöglichkeiten zwischen „nicht wichtig“ (1) und „sehr wichtig“ (5) wurde im Durchschnitt eine 2,0 bewertet. Daraus folgt, dass die Elemente im Film, die Emotionen ausdrücken, nicht zu sehr in den Vordergrund der Audiodeskription gestellt werden sollten. Außerdem bestärkt es erneut die Feststellung, dass eine Audiodeskription nicht existenziell wichtig für die Emotionsvermittlung ist.

14. Auf welche Aspekte soll bei der ausführlichen Beschreibung eingegangen werden?

Die Teilnehmenden waren sich einig, dass zuallererst die für das Verständnis der Handlung relevanten Informationen beschrieben werden müssen. Darüber hinaus sind die Interessen vielfältig und subjektiv. Manche hätten gerne mehr Informationen zur Umgebung, andere lieber eine genauere Beschreibung der Mimik und Gestik. Ein Proband interessiert sich für die

Kameraführung und den Schnitt, da er sich mit diesen Aspekten in seinen sehenden Zeiten beschäftigt hat.

Innerhalb meiner Stichprobe ist die Beschreibung von Farben in der Wichtigkeit weiter hinten einzuordnen. Insbesondere die geburtsblinden Testpersonen legen wenig Wert darauf.

Allgemeine Fragen zum System:

15. *Können Sie sich vorstellen, dass das System der personalisierten Audiodeskription mit Sprachsynthese bei 90–120-minütigen Filmen funktioniert?*

Diese Frage wurde nahezu einstimmig von den Testpersonen bejaht, allerdings mit der Ergänzung, dass es schwierig einzuschätzen ist und eine Testung diesbezüglich sinnvoll wäre.

Ein Proband teilt seine Befürchtung mit, dass mitschauende sehende Menschen von der künstlichen Stimme und dem Wechseln der Geschwindigkeiten noch schneller genervt sein könnten, als von der standardisierten Audiodeskription.

16. *Wenn es möglich wäre, den personalisierten Ansatz mit gleichem Zeitaufwand und den gleichen Produktionskosten wie eine normale, menschliche Audiodeskription herzustellen, welches System würden Sie dann bevorzugen? Die menschliche Audiodeskription oder den vorgestellten Ansatz?*

Eindeutig ist, dass die Antwort viel facettenreicher ist und nicht schlicht mit einer Entscheidungsfrage beantwortet werden kann. Die Vielschichtigkeit wurde mit den Fragen zuvor aufgezeigt und in der Auswertung dieser ausführlich berücksichtigt.

Die Beantwortung der Frage 16 sollte jedoch nicht außen vorgelassen werden, daher stellt die Grafik in Abbildung 24 die Ergebnisse dar.

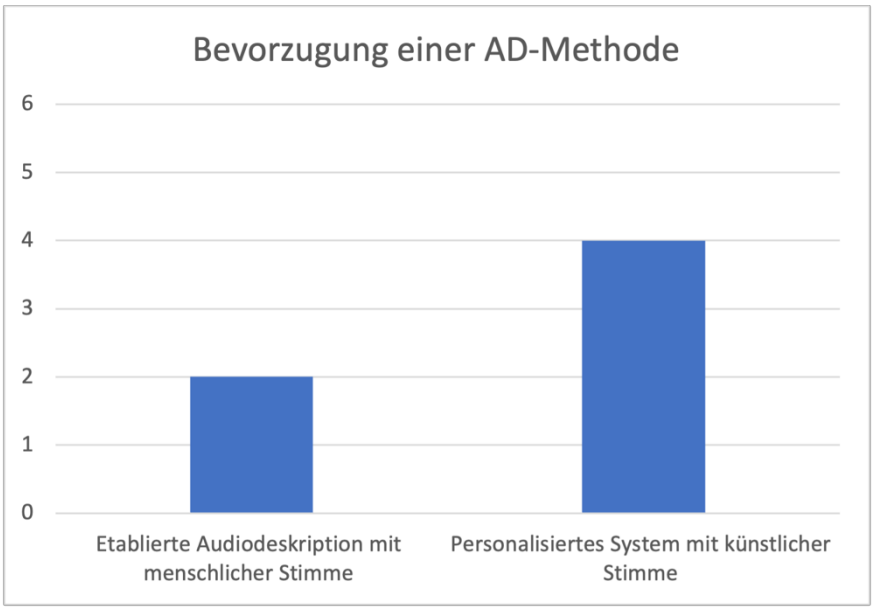

Abbildung 24: Bevorzugung der personalisierten AD mit künstlicher Stimme oder der AD mit menschlicher Stimme (eigene Darstellung)

Mit vier Stimmen für das personalisierte System und zwei Stimmen für die etablierte Audiodeskription besteht innerhalb der hier einbezogenen Stichprobe eine leichte Tendenz zur Bevorzugung des personalisierten Systems. Die Hauptbegründungen der befürwortenden Personen ist das Erreichen einer größeren Zielgruppe durch die Individualisierungsoptionen dieses Systems. Außerdem erscheint die Möglichkeit, eine gesteigerte Inhaltsmenge zu erhalten, äußert attraktiv. Die Einwände der Personen, die anders entschieden haben, basieren darauf, dass eine menschliche Stimme den Vorteil besitzt, die Geschwindigkeiten bereits flexibel innerhalb der Audiodeskription des Films gestalten zu können, sowie dass die menschliche Stimme „einfach schöner klingt".

Die Konsequenz, dass die Arbeit für einige Sprecherinnen und Sprecher durch den Einsatz von künstlichen Stimmen reduziert werden könnte, hat die Entscheidung ebenfalls beeinflusst.

17. Haben Sie Verbesserungsvorschläge für das System?

Bezüglich der Geschwindigkeit, der Genres und des Inhalts wurden keine weiteren Verbesserungen vorgeschlagen. Angemerkt wurde, dass Tastenkombinationen zum Wechseln der AD-Versionen angenehm wären und dass darüber nachzudenken ist, wie ein solches System bei einem Fernseher oder einem DVD-Player verwendet werden könnte.

6.5 Resumé und Fazit

Die dargestellten Ergebnisse melden eine positive Reaktion mit Einschränkungen auf das neuartige System zurück.

Die Qualität der künstlichen Stimme wurde mit einer Punktzahl von 4,3 von 5 Punkten als gut bewertet, welches die Akzeptanz synthetisierter Stimmen bestärkt. Die Emotionsvermittlung wurde durch das System mit künstlicher Stimme nur minimal eingeschränkt, was durch ein klares Meinungsbild bezüglich der Ansicht, dass die Emotionsvermittlung nicht Aufgabe der Audiodeskription ist, resultiert.

Als bevorzugte Geschwindigkeit wurde die schnelle Stufe mit 58% gewählt, wodurch sich schlussfolgern lässt, dass ein Bedürfnis nach einer erhöhten Inhaltsvermittlung existiert. Im Gegensatz dazu wurde von keiner Person die langsame Geschwindigkeit präferiert, jedoch zum Teil für sinnvoll gehalten. Das Erstellen von drei verschiedenen AD-Skripten, zwischen denen einen Wechsel ohne Logiklücken möglich ist, stellt sicherlich eine Herausforderung dar. Allerdings wurde diese Wechselfunktion von fünf von sechs Teilnehmenden als sehr sinnvoll empfunden.

Eine besonders gute Eignung dieses Systems hinsichtlich eines der getesteten Genres konnte nicht festgestellt werden, sondern lediglich eine Tendenz zum „Actionfilm“ hat sich herauskristallisiert. Daraus lässt sich schließen, dass das personalisierte System für jedes der drei Genres eine erfolgversprechende Möglichkeit bietet.

Zwei von sechs Personen haben aufgrund der in sich flexiblen, angenehm klingenden menschlichen Stimme das etablierte System bevorzugt, wohingegen die anderen vier Testpersonen bei einem gleichen Kosten- und Zeitaufwand der beiden Methoden, für die Einführung des neuartigen Systems stimmten. Argumentiert wurde, dass das Erreichen einer größeren Zielgruppe durch diese Personalisierung einen erheblichen Mehrwert bietet. Nun stellt sich folglich die Frage, ob die Umsetzung des personalisierten Systems mit gleichem Kosten- und Zeitaufwand wie der von der bisherigen Audiodeskription, erreicht werden kann. Dafür spricht das Wegfallen der Kosten für die Sprachaufnahme und des dafür gebuchten Tonstudios. Dagegen sprechen die Herstellungskosten von drei Audiodeskriptions-Skripten, wobei der Aufwand davon sicherlich nicht das Dreifache beträgt.

Solche und andere Gedanken, die im nächsten Kapitel aufgeführt werden, müssten vor einer Einführung dieses Systems berücksichtigt und durchdacht werden.

7 Ausblick

Trotz der vielversprechenden Ergebnisse des Usability-Tests müssten weitere Untersuchungen vorgenommen werden, um einen absoluten Mehrwert des Systems feststellen zu können.

Wie bereits erwähnt wäre eine ausführliche Kostenanalyse notwendig. Außerdem müsste der Ansatz an einer größeren Studienpopulation getestet werden, um herausfinden zu können, ob die langsame Geschwindigkeitsstufe Gebrauch finden würde oder nicht. Die Population sollte daher aus verschiedenen Bildungsschichten bestehen und ebenfalls alte Menschen beinhalten, die Audiodeskriptionen verwenden, obwohl sie nicht unbedingt blind sind.

Zudem wäre es interessant, statt einer Naturdokumentation beispielsweise eine technische Dokumentation oder ein Lehrvideo zu testen, da alle getesteten Filmausschnitte viel emotionalen Inhalt vorzuweisen hatten.

Wichtig ist ebenfalls eine Testung des Systems anhand von Filmen mit kompletter Länge. Dadurch ließe sich herausfinden, ob die Eintönigkeit der künstlichen Stimme die Konzentration beeinflusst und ob das Wechseln der Ausführlichkeitsstufen Logiklücken aufweist. Beim Eintreten dieses Falls wäre die personalisierte Audiodeskription mit künstlicher Stimme möglicherweise eine Lösung für Social-Media-Plattformen und anderen Online-Diensten, die kürzere Videos zeigen. Andernfalls müsste man sich überlegen, wie sich die Bedienung dieses Systems nicht nur im Internet, sondern auch für Filme im Fernsehen gestalten ließe.

Die Durchführung dieser Studie hat sehr viel Freude bereitet und verleitet zu weiteren Gedankenexperimenten. Diese beziehen sich vor allem auf weitere Personalisierungsoptionen.

Die verschiedenen Versionen der Audiodeskription könnten sich nicht nur auf die Inhaltsmenge beziehen, sondern auch auf die Art und Weise der Beschreibung. Man könnte beispielsweise eine Unterscheidung zwischen subjektiver und objektiver Beschreibung einführen, sowie die Texte auf Geburtsblinde oder Späterblindete anpassen. Außerdem würde die in dieser Arbeit aufgeführte Technologie der automatischen Skripterstellung für den hier untersuchten Ansatz einen enormen Vorteil bieten, da die Hauptarbeit hierbei aus der Erstellung der drei Skripte besteht.

Fakt ist, dass eine Personalisierung in allen Bereichen von Multimedia stattfindet. Die Resultate dieser Studie belegen, dass ebenfalls eine Nachfrage im Bereich der Audiodeskription besteht und ein personalisiertes System, wie es hier vorgeschlagen wird, bei einigen AD-Nutzenden auf Interesse und Begeisterung stößt.

8 Literaturverzeichnis

„Able Player Demos“. Zugegriffen 12. März 2023. https://ableplayer.github.io/ableplayer/demos/.

„ADLAB Audio Description guideline“. Zugegriffen 13. Januar 2023. http://www.adlabproject.eu/Docs/adlab%20book/#index.

Anishchenko, Alla. „Emotionstransfer bei der Audiodeskription“. *Linguistische Treffen in Wroc\law*, Nr. 18 (2020): 23–32.

„ARD Mediathek“. Zugegriffen 24. März 2023. https://www.ardmediathek.de/video/erlebnis-erde/wildes-kalifornien-1-stroeme-des-lebens-hoerfassung/das-erste/Y3JpZDovL2Rhc2Vyc3RlLmRlL2VybGVibml-zIGVyZGUvMjAyMyowMS0wOV8yMCoxNS1NRVovYXVkaW9kZXNrcmlwdGlvbg

„ARD Mediathek“. Zugegriffen 24. März 2023. https://www.ardmediathek.de/video/in-aller-freundschaft/zu-endstoff-597-hoerfassung/mdr-fernsehen/Y3JpZDovL21kci5kZS9iZWlocmFnL2Ntcy80Nzg4YzRhMC1hND-MwLTQ2MDYtOGZiZC1mMjU5NzU1MzY2MWYvYXVkaW9kZXNrcmlwdGlvbg

Arik, Sercan, Jitong Chen, Kainan Peng, Wei Ping, und Yanqi Zhou. „Neural Voice Cloning with a Few Samples“. In *Advances in Neural Information Processing Systems*, Bd. 31. Curran Associates, Inc., 2018. https://proceedings.neurips.cc/paper/2018/hash/4559912e7a94a9c32b09d894f2bc3c82-Abstract.html.

„Audio demos“. Zugegriffen 1. März 2023. https://audiodemos.github.io/.

Bäckström, Tom. *Speech Coding*. Signals and Communication Technology. Cham: Springer International Publishing, 2017. https://doi.org/10.1007/978-3-319-50204-5.

Barrierefreiheit | Schulung, Begleitung und Tests. „Barrierefreiheit von YouTube“, 16. März 2018. https://www.netz-barrierefrei.de/wordpress/barrierefreies-web-2-0-ein-leitfaden-zu-social-media-und-behinderung/barrierefreiheit-von-youtube/.

Benecke, Bernd. „Audiodeskription - Methoden und Techniken der Filmbeschreibung“. In *Handbuch Barrierefreie Kommunikation*, herausgegeben von Christiane Maass und Isabel Rink, 455–70. Kommunikation - Partizipation - Inklusion, Band 3. Berlin: Frank & Timme, Verlag für wissenschaftliche Literatur, 2019.

———. *Audiodeskription als partielle Translation : Modell und Methode*. mitSprache. Berlin ; Münster: LIT, 2014.

Bernabé, Rocío, und Pilar Orero. „Easier audio description: Exploring the potential of Easy-to-Read principles in simplifying AD“. In *Innovation in Audio Description Research*. Routledge, 2020.

Beuth, Patrick. „Lyrebird ausprobiert: Diese Software imitiert jede Stimme“. *Der Spiegel*, 28. Januar 2018, Abschn. Netzwelt. https://www.spiegel.de/netzwelt/web/lyrebird-diese-software-imitiert-jede-stimme-a-1189829.html.

„BITV 2.0 - Verordnung zur Schaffung barrierefreier Informationstechnik nach dem Behindertengleichstellungsgesetz“. Zugegriffen 22. März 2023. https://www.gesetze-im-internet.de/bitv_2_0/BJNR184300011.html.

Bogucki, Łukasz, und Mikołaj Deckert. *The Palgrave Handbook of Audiovisual Translation and Media Accessibility*. 1st ed. 2020. Cham: Springer International Publishing, 2020.

Böhm, Markus. „(S+) Stimmengeneratoren und Audio-Fakes: Online-Trolle lassen Emma Watson »Mein Kampf« vorlesen“. *Der Spiegel*, 31. Januar 2023, Abschn. Netzwelt. https://www.spiegel.de/netzwelt/web/elevenlabs-stimmengenerator-online-trolle-lassen-emma-watson-mein-kampf-vorlesen-a-780f1457-5a1c-40e0-b909-57835f89125d.

Braun, Sabine. „Creating coherence in audio description“. *Meta: Journal des traducteurs/Meta: Translators' Journal* 56, Nr. 3 (2011): 645–62.

Braun, Sabine, und Kim Starr. „Automating audio description“. In *The Routledge Handbook of Audio Description*, herausgegeben von Christopher Taylor und Elisa Perego, 391–406. London: Routledge, 2022. https://doi.org/10.4324/9781003003052.

———. „Finding the Right Words: Investigating Machine-Generated Video Description Quality Using a Corpus-Based Approach“. *Journal of Audiovisual Translation* 2, Nr. 2 (31. Dezember 2019): 11–35. https://doi.org/10.47476/jat.v2i2.103.

Brescia-Zapata, Marta. „The Present and Future of Accessibility Services in VR360 Players.“ *InTRAlinea: Online Translation Journal* 24 (2022).

Buchkremer, Rüdiger. „Natural Language Processing in der KI“. In *Künstliche Intelligenz in Wirtschaft & Gesellschaft: Auswirkungen, Herausforderungen & Handlungsempfehlungen*, herausgegeben von Rüdiger Buchkremer, Thomas Heupel, und Oliver Koch, 29–45. FOM-Edition. Wiesbaden: Springer Fachmedien, 2020. https://doi.org/10.1007/978-3-658-29550-9_2.

Campos, Virginia P., Tiago M. U. de Araújo, Guido L. de Souza Filho, und Luiz M. G. Gonçalves. „CineAD: A System for Automated Audio Description Script Generation for the Visually Impaired“. *Universal Access in the Information Society* 19, Nr. 1 (1. März 2020): 99–111. https://doi.org/10.1007/s10209-018-0634-4.

Caro, Marina Ramos. „Testing audio narration: the emotional impact of language in audio description“. *Perspectives* 24, Nr. 4 (2016): 606–34.

Carré, René, Pierre Divenyi, und Mohamad Mrayati. *Speech: A Dynamic Process*. Berlin ; Boston: de Gruyter, 2017.

Chen, Kuo, und Xuebin Sun. „CRCTTS: Convolution-Recurrent-Convolution Text-to-Speech System“. In *Proceedings of the 2022 2nd International Conference on Control and*

Intelligent Robotics, 774–77. ICCIR '22. New York, NY, USA: Association for Computing Machinery, 2022. https://doi.org/10.1145/3548608.3559304.

Chmiel, Agnieszka, und Iwona Mazur. „A homogenous or heterogeneous audience? Audio description preferences of persons with congenital blindness, non-congenital blindness and low vision". *Perspectives* 30, Nr. 3 (2022): 552–67.

Christian Schiffer, Bayerischer Rundfunk. „#failoftheweek: Wie Fake-Audios das Netz fluten – und dabei täuschend echt wirken", 2. Februar 2023. https://www.br.de/radio/bayern2/sendungen/zuendfunk/fail-of-the-week-warum-fake-audios-und-voice-cloning-die-neuen-deepfakes-sind-100.html.

Chrome Web-Store. „Speak Subtitles for Youtube". Chrome-Erweiterungen, 22. Juni 2022. https://chrome.google.com/webstore/detail/speak-subtitles-for-youtu/fjoiihoancoimepbgfcmopaciegpigpa.

Evans, Michael, Lianne Kerlin, Joanne Parkes, und Todd Burlington. „"I want to be independent. I want to make informed choices.": An Exploratory Interview Study of the Effects of Personalisation of Digital Media Services on the Fulfilment of Human Values". In *ACM International Conference on Interactive Media Experiences*, 325–30. IMX '22. New York, NY, USA: Association for Computing Machinery, 2022. https://doi.org/10.1145/3505284.3532977.

Fellbaum, Klaus. *Sprachverarbeitung und Sprachübertragung*. Berlin, Heidelberg: Springer, 2012. https://doi.org/10.1007/978-3-642-31503-9.

Fix, Ulla. *Hörfilm : Bildkompensation durch Sprache ; linguistisch-filmisch-semiotische Untersuchungen zur Leistung der Audiodeskription in Hörfilmen am Beispiel des Films „Laura, mein Engel" aus der „Tatort"-Reihe*. Philologische Studien und Quellen. Berlin: Schmidt, 2005.

Fryer, Louise. *An Introduction to Audio Description : A Pratical Guide*. London ; New York: Taylor & Francis Ltd, 2016.

Gold, Bernard, Daniel Patrick Whittlesey Ellis, und Nelson Morgan. *Speech and Audio Signal Processing : Processing and Perception of Speech and Music*. 2nd ed. Oxford: John Wiley & Sons, 2011.

Gzara, Noura. „Der Erstellungsprozess der Audiodeskription". *Audiodeskription verständlich erklärt*, 2021, 54.

Hämmer, Karin. „Personenkennzeichnungen in Audiodeskriptionen". In *Hörfilm : Bildkompensation durch Sprache ; linguistisch-filmisch-semiotische Untersuchungen zur Leistung der Audiodeskription in Hörfilmen am Beispiel des Films „Laura, mein Engel" aus*

der „Tatort"-Reihe, herausgegeben von Ulla Fix, 87–98. Philologische Studien und Quellen. Berlin: Schmidt, 2005.

Heerdegen-Wessel, Uschi. „Barrierefreie Angebote des NDR und der ARD - Stand, Aufgaben, Ziele". In *Handbuch Barrierefreie Kommunikation*, herausgegeben von Christiane Maass und Isabel Rink, 725–40. Kommunikation - Partizipation - Inklusion, Band 3. Berlin: Frank & Timme, Verlag für wissenschaftliche Literatur, 2019.

Hermosa-Ramírez, Irene. „Profiling audio description service providers". In *The Routledge Handbook of Audio Description*, 295–311. Routledge, 2022.

Hirvonen, Maija Inkeri, und Reinhold Schmitt. „Blindheit als Ressource: Zur professionellen Kompetenz eines blinden Teammitglieds bei der gemeinsamen Anfertigung einer Audiodeskription". *Gesprächsforschung*, 2018.

„Hoerfilm e.V. | Audiodeskription". Zugegriffen 6. Dezember 2022. https://hoerfilmev.de/audiodeskription/.

Jaun, René. „Hörfilme für Blinde: Wenn die Computerstimme den Film beschreibt", 24. Februar 2021. https://www.netzwoche.ch/news/2021-02-24/hoerfilme-fuer-blinde-wenn-die-computerstimme-den-film-beschreibt.

Jimenez Hurtado, Catalina, und Silvia Soler Gallego. „Multimodality, Translation and Accessibility: A Corpus-Based Study of Audio Description". *Perspectives* 21, Nr. 4 (Dezember 2013): 577–94. https://doi.org/10.1080/0907676X.2013.831921.

Jüngst, Heike Elisabeth. *Audiovisuelles Übersetzen : ein Lehr- und Arbeitsbuch*. 2., Überarbeitete und Erweiterte Auflage. Tübingen: Narr Francke Attempto, 2020.

Kordijazi, Amir, Tian Zhao, Jun Zhang, Khaled Alrfou, und Pradeep Rohatgi. „A Review of Application of Machine Learning in Design, Synthesis, and Characterization of Metal Matrix Composites: Current Status and Emerging Applications". *JOM* 73, Nr. 7 (Juli 2021): 2060–74. https://doi.org/10.1007/s11837-021-04701-2.

Kuniavsky, Mike. *Observing the User Experience*. 1st edition. Morgan Kaufmann, 2003.

Kurch, Alexander. „Produktionsprozesse der Hörgeschädigten-Untertitelungen und Audiodeskription: Potenziale teilautomatisierter Prozessbeschleunigung mittels (Sprach-)Technologien". In *Handbuch Barrierefreie Kommunikation*, herausgegeben von Christiane Maass und Isabel Rink, 437–54. Kommunikation - Partizipation - Inklusion, Band 3. Berlin: Frank & Timme, Verlag für wissenschaftliche Literatur, 2019.

Leegaard, Jack, Jan Østergaard, Søren Holdt Jensen, und Ram Zamir. „Practical Design of Delta-Sigma Multiple Description Audio Coding". *EURASIP Journal on Audio, Speech, and Music Processing* 2014, Nr. 1 (22. April 2014): 16. https://doi.org/10.1186/1687-4722-2014-16.

Maass, Christiane, und Isabel Rink, Hrsg. *Handbuch Barrierefreie Kommunikation*. Kommunikation - Partizipation - Inklusion, Band 3. Berlin: Frank & Timme, Verlag für wissenschaftliche Literatur, 2019.

Mark8-36. „Emma Watson Reads Mein kampf AI Generated“, 8. Februar 2023. https://www.youtube.com/watch?v=HkwVxyygbdo.

Matamala, Anna, und Pilar Orero, Hrsg. *Researching Audio Description*. London: Palgrave Macmillan UK, 2016. https://doi.org/10.1057/978-1-137-56917-2.

Mazur, Iwona. „A functional approach to audio description“. *Journal of Audiovisual Translation* 3, Nr. 2 (2020): 226–45.

Mazur, Iwona, und Agnieszka Chmiel. „A homogenous or heterogeneous audience? Audio description preferences of persons with congenital blindness, non-congenital blindness and low vision | 10.1080/0907676x.2021.1913198“, 2021. https://sci-hub.ee/https:/www.tandfonline.com/doi/abs/10.1080/0907676X.2021.1913198.

———. „AD reception research: Some methodological considerations“, 2011.

———. „Audio description made to measure: Reflections on interpretation in AD based on the Pear Tree Project data“. In *Audiovisual translation and media accessibility at the crossroads*, 173–88. Brill, 2012.

Microsoft. „Benutzerdefinierte neuronale Stimme in der Übersicht: Speech-Dienst - Azure Cognitive Services“, 3. November 2022. https://learn.microsoft.com/de-de/azure/cognitive-services/speech-service/custom-neural-voice.

Minutella, Vincenza. „Audiodescription Software Tools“. In *The Routledge Handbook of Audio Description*, herausgegeben von Christopher Taylor und Elisa Perego, 331–52. London: Routledge, 2022. https://doi.org/10.4324/9781003003052.

Moreno, Lourdes, María González-García, Paloma Martínez, und Yolanda González. „Checklist for Accessible Media Player Evaluation“. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility*, 367–68. Baltimore Maryland USA: ACM, 2017. https://doi.org/10.1145/3132525.3134791.

Morgner, Henrike, und Steffen Pappert. „Einstellungsanalyse und Transkription ‚Laura, mein Engel‘“. In *Hörfilm : Bildkompensation durch Sprache ; linguistisch-filmisch-semiotische Untersuchungen zur Leistung der Audiodeskription in Hörfilmen am Beispiel des Films „Laura, mein Engel“ aus der „Tatort“-Reihe*, herausgegeben von Ulla Fix, V. Philologische Studien und Quellen. Berlin: Schmidt, 2005.

Morgner, Henrike, und Steffen Pappert. „Darstellung des Untersuchungsmaterials: Sequenzprotokoll, Einstellungsanalyse und Transkription“. In *Hörfilm : Bildkompensation durch Sprache ; linguistisch-filmisch-semiotische Untersuchungen zur Leistung der*

Audiodeskription in Hörfilmen am Beispiel des Films „Laura, mein Engel" aus der „Tatort"-Reihe, herausgegeben von Ulla Fix, 13–32. Philologische Studien und Quellen. Berlin: Schmidt, 2005.

Nakajima, Sawako, und Kazutaka Mitobe. „Novel Software for Producing Audio Description Based on Speech Synthesis Enables Cost Reduction without Sacrificing Quality". *Universal Access in the Information Society* 21, Nr. 2 (1. Juni 2022): 405–18. https://doi.org/10.1007/s10209-022-00873-z.

NDR. „Audio description guidelines". Zugegriffen 12. Januar 2023. https://www.ndr.de/fernsehen/barrierefreie_angebote/audiodeskription/Audio-description-guidelines,audiodeskription142.html.

„Netflix". Zugegriffen 24. März 2023. https://www.netflix.com/browse/audio-description.

Neto, José Monserrat, André P. Freire, Sabrina S. Souto, und Ramon S. Abílio. „Usability Evaluation of a Web System for Spatially Oriented Audio Descriptions of Images Addressed to Visually Impaired People". In *Universal Access in Human-Computer Interaction. Universal Access to Information and Knowledge*, herausgegeben von Constantine Stephanidis und Margherita Antona, 154–65. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2014. https://doi.org/10.1007/978-3-319-07440-5_15.

Orero, Pilar. „Audio description personalisation". In *The Routledge Handbook of Audio Description*, 407–19. Routledge, 2022.

———. „Audio description personalisation". In *The Routledge Handbook of Audio Description*. Routledge, 2022.

———. „Audio description personalisation". In *The Routledge Handbook of Audio Description*, herausgegeben von Christopher Taylor und Elisa Perego, 407–19. London: Routledge, 2022. https://doi.org/10.4324/9781003003052.

AccessibilityOz. „OzPlayer". Zugegriffen 12. März 2023. https://www.accessibilityoz.com/oz-player/.

Pantula, Muralidhar, und K. S. Kuppusamy. „AuDIVA: A Tool for Embedding Audio Descriptions to Enhance Video Accessibility for Persons with Visual Impairments". *Multimedia Tools and Applications* 78, Nr. 14 (1. Juli 2019): 20005–18. https://doi.org/10.1007/s11042-019-7363-4.

Papachristos, Nikiforos M., Ioannis Vrellis, und Tassos A. Mikropoulos. „A Comparison between Oculus Rift and a Low-Cost Smartphone VR Headset: Immersive User Experience and Learning". In *2017 IEEE 17th International Conference on Advanced Learning Technologies (ICALT)*, 477–81. Timisoara, Romania: IEEE, 2017. https://doi.org/10.1109/ICALT.2017.145.

Pfister, Beat, und Tobias Kaufmann. *Sprachverarbeitung : Grundlagen und Methoden der Sprachsynthese und Spracherkennung*. 2. Aufl. 2017. Berlin, Heidelberg: Springer Vieweg, 2017.

———. *Sprachverarbeitung : Grundlagen und Methoden der Sprachsynthese und Spracherkennung*. 2. Aufl. 2017. Berlin, Heidelberg: Springer Vieweg, 2017.

Poethe, Hannelore. „Audiodeskription - Enstehung und Wesen einer Textsorte". In *Hörfilm : Bildkompensation durch Sprache ; linguistisch-filmisch-semiotische Untersuchungen zur Leistung der Audiodeskription in Hörfilmen am Beispiel des Films „Laura, mein Engel" aus der „Tatort"-Reihe*, herausgegeben von Ulla Fix, 33–48. Philologische Studien und Quellen. Berlin: Schmidt, 2005.

Prasanna, S. R. Mahadeva. *Speech and Computer : 24th International Conference, SPECOM 2022, Gurugram, India, November 14–16, 2022, Proceedings*. 1st ed. 2022. Lecture Notes in Artificial Intelligence. Cham: Springer International Publishing, 2022.

„Prüfschritte BITV-Test / EN 301 549 (Web) | BIK BITV-Test Ergebnisse und Methodik | BIK BITV-Test". Zugegriffen 20. März 2023. https://ergebnis.bitvtest.de/pruefverfahren/bitv-20-web.

Raemont, Nina. „Here's How to Use YouTube's New Multilanguage Audio Feature". CNET, 27. Februar 2023. https://www.cnet.com/tech/services-and-software/heres-how-to-use-youtubes-new-multilanguage-audio-feature/.

Ramos, Marina. „The emotional experience of films: Does audio description make a difference?" *The Translator* 21, Nr. 1 (2015): 68–94.

„Richtlinien für barrierefreie Webinhalte (WCAG) 2.1". Zugegriffen 6. März 2023. https://outline-rocks.github.io/wcag/translations/CAT-WCAG21-DE-20211004/#dfn-erweiterte-audiodeskription.

Rocha Façanha, Agebson, Adonias Caetano de Oliveira, Marcos Vinicius de Andrade Lima, Windson Viana, und Jaime Sánchez. „Audio Description of Videos for People with Visual Disabilities". In *Universal Access in Human-Computer Interaction. Users and Context Diversity*, herausgegeben von Margherita Antona und Constantine Stephanidis, 505–15. Lecture Notes in Computer Science. Cham: Springer International Publishing, 2016. https://doi.org/10.1007/978-3-319-40238-3_48.

Sackl, Andreas, Franziska Graf, Raimund Schatz, und Manfred Tscheligi. „Ensuring Accessibility: Individual Video Playback Enhancements for Low Vision Users". In *The 22nd International ACM SIGACCESS Conference on Computers and Accessibility*, 1–4. Virtual Event Greece: ACM, 2020. https://doi.org/10.1145/3373625.3417997.

Sade, Jack, Komal Naz, und Malgorzata Plaza. „Enhancing Audio Description: A Value Added Approach". In *Computers Helping People with Special Needs*, herausgegeben von Klaus

Miesenberger, Arthur Karshmer, Petr Penaz, und Wolfgang Zagler, 270–77. Lecture Notes in Computer Science. Berlin, Heidelberg: Springer, 2012. https://doi.org/10.1007/978-3-642-31522-0_40.

Salway, Andrew. „A corpus-based analysis of audio description“. In *Media for all*, 151–74. Brill, 2007.

Schneider, Remo. „Development and validation of a concept for layered audio descriptions“. Master-Thesis, Hochschule der Medien, 2021.

Schruhl, Stefanie. „Audiodeskription in der Praxis - Bericht einer blinden Rezipientin und Hörfilmautorin“. In *Handbuch Barrierefreie Kommunikation*, herausgegeben von Christiane Maass und Isabel Rink, 771–80. Kommunikation - Partizipation - Inklusion, Band 3. Berlin: Frank & Timme, Verlag für wissenschaftliche Literatur, 2019.

Snyder, Joel. „Audio description: The visual made verbal“. In *International Congress Series*, 1282:935–39. Elsevier, 2005.

„Speech Studio - Microsoft Azure“. Zugegriffen 21. März 2023. https://speech.microsoft.com/audiocontentcreation.

Szarkowska, Agnieszka. „Text-to-speech audio description: towards wider availability of AD“. *The Journal of Specialised Translation* 15, Nr. 1 (2011): 142–62.

Taylor, Christopher, und Elisa Perego, Hrsg. *The Routledge Handbook of Audio Description*. London: Routledge, 2022. https://doi.org/10.4324/9781003003052.

Tyfour, Maher. *Sprachmacht auf engstem Raum: die Inszenierung der Stadt in den Hörfilmen der Münchner Tatort-Filmserie: eine korpusgeleitete Studie zurAudiodeskription*. Easy-plain-accessible, vol. 7. Berlin: Frank & Timme, Verlag für wissenschaftliche Literatur, 2021.

„Übersicht — Personalisierte Audiodeskription“. Zugegriffen 15. März 2023. https://franzi.bf-lernen.de/.

Viswanathan, Lakshmie Narayan, Troy McDaniel, und Sethuraman Panchanathan. „Audio-Haptic Description in Movies“. In *HCI International 2011 – Posters' Extended Abstracts*, herausgegeben von Constantine Stephanidis, 414–18. Communications in Computer and Information Science. Berlin, Heidelberg: Springer, 2011. https://doi.org/10.1007/978-3-642-22098-2_83.

Walczak, Agnieszka. „Audio Description on Smartphones: Making Cinema Accessible for Visually Impaired Audiences“. *Universal Access in the Information Society* 17, Nr. 4 (November 2018): 833–40. https://doi.org/10.1007/s10209-017-0568-2.

———. „Audio Description on Smartphones: Making Cinema Accessible for Visually Impaired Audiences“. *Universal Access in the Information Society* 17, Nr. 4 (1. November 2018): 833–40. https://doi.org/10.1007/s10209-017-0568-2.

Walczak, Agnieszka, und Louise Fryer. „Creative description: The impact of audio description style on presence in visually impaired audiences“. *British Journal of Visual Impairment* 35, Nr. 1 (2017): 6–17.

Wang, Yujia, Wei Liang, Haikun Huang, Yongqi Zhang, Dingzeyu Li, und Lap-Fai Yu. „Toward Automatic Audio Description Generation for Accessible Videos“. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–12. Yokohama Japan: ACM, 2021. https://doi.org/10.1145/3411764.3445347.

Yos, Gabriele. „Verknüpfungen von Audiodeskription und Filmdialog“. In *Hörfilm : Bildkompensation durch Sprache ; linguistisch-filmisch-semiotische Untersuchungen zur Leistung der Audiodeskription in Hörfilmen am Beispiel des Films „Laura, mein Engel“ aus der „Tatort“-Reihe*, herausgegeben von Ulla Fix, 99–116. Philologische Studien und Quellen. Berlin: Schmidt, 2005.

„YouTube“. Zugegriffen 12. März 2023. https://www.youtube.com/.

„Zahlen & Fakten zu Blindheit und Sehbehinderung“. Zugegriffen 16. März 2023. https://www.dbsv.org/zahlen-fakten.html.

„ZDF Mediathek“. Zugegriffen 24. März 2023. https://www.zdf.de/uri/6b9eb468-380b-49b9-8d60-7bcc63d9559e.

Zeeck, Achim. *Speech Application SDK Mit ASP.NET*, 2005. https://link.springer.com/book/10.1007/3-540-28086-3.